CSEF

Centre for Studies in Economics and Finance

WORKING PAPER NO. 278

Imputation of Missing Data in Waves 1 and 2 of SHARE

Dimitris Christelis

March 2011

University of Naples Federico II

University of Salerno

Bocconi University, Milan

CSEF - Centre for Studies in Economics and Finance
DEPARTMENT OF ECONOMICS – UNIVERSITY OF NAPLES
80126 NAPLES - ITALY
Tel. and fax +39 081 675372 – e-mail: csef@unisa.it

CSEF
Centre for Studies in Economics and Finance

WORKING PAPER NO. 278

Imputation of Missing Data in Waves 1 and 2 of SHARE

Dimitris Christelis*

Abstract

The Survey of Health, Ageing and Retirement in Europe (SHARE), like all large household surveys, suffers from the problem of item non-response, and hence the need of imputation of missing values arises. In this paper I describe the imputation methodology used in the first two waves of SHARE, which is the fully conditional specification approach of van Buuren, Brand, Groothuis-Oudshoorn, and Rubin (2006). Methods for assessing the convergence of the imputation process are also discussed. Finally, I give details on numerous issues affecting the implementation of the imputation process that are particular to SHARE.

JEL Classification: C81, C83

Keywords: Missing Data; Multiple Imputation; Markov Chain Monte Carlo; SHARE.

Acknowledgements I am grateful to Guglielmo Weber for his encouragement, and for all the input that he has given me in the course of many discussions. I would also like to thank Arthur Kenickell for providing inspiration, as well as many useful tips on how to perform multiple imputation. Valuable contributions to the SHARE imputation project have also been made by Viola Angelini, Agar Brugiavini, Lisa Callegaro, Danilo Cavapozzi, Enrica Croda, Loretti Dobrescu, Thomas Georgiadis, Anne Laferrère, and Omar Paccagnella. SHARE data collection in 2004-2007 was primarily funded by the European Commission through its 5th and 6th framework programs (project numbers QLK6-CT-2001- 00360; RII-CT- 2006-062193; CIT5-CT-2005-028857). Additional funding by the US National Institute on Aging (grant numbers U01 AG09740-13S2; P01 AG005842; P01 AG08291; P30 AG12815; Y1-AG-4553-01; OGHA 04-064; R21 AG025169) as well as by various national sources is gratefully acknowledged (see http://www.share-project.org for a full list of funding institutions).

* SHARE, CSEF and CFS. E-mail address: dimitris [dot] christelis [at] gmail [dot] com

Table of contents

I. Introduction

The Survey of Health, Ageing and Retirement in Europe (SHARE), like all large household surveys, suffers from the problem of item non-response. There are many reasons why this is the case, including the length of the questionnaire, respondents' privacy concerns, physical and mental health problems, cognitive limitations, and their lack of free time due to work obligations, or to the provision of care to young children or elderly relatives.

One way to deal with the problem of missing data would be to fill in the missing values as much as possible using information available from other sources (e.g. the remarks made by survey interviewers), but then leave the remaining missing values as they are. As a result, the users of the data would make their own decisions on how to deal with the missing data. This would almost surely imply that many of them would analyze the data after discarding all observations with missing values. This decision might not even be taken by the users themselves, but rather by the statistical software that they are using, given that, as a rule, the latter will discard all observations with missing data before producing the results asked for.

While the decision to not use any observations with missing values might superficially appear to lead to a "clean" analysis of the data, in reality it implies making the strongest possible assumption about them, namely that the observations containing missing values are not in any way different from those without missing values. If this were true, then the part of the sample that would be left after deleting all observations with missing values would still be representative of the original sample. Essentially, this assumption implies that all missingness is completely random, i.e., that the mechanism that generates missing data is uncorrelated with any variables that may or may not be present in the survey. This assumption is, however, almost surely violated: as already discussed, there are many reasons that can lead to item non-response, which thus becomes non-random. A violation of the missing completely at random (MCAR) assumption will likely make analyses based only on observations with complete records biased and inconsistent (Rubin, 1987; Little and Rubin, 2002).

In addition, given the prevalence of missing data typically encountered in large household surveys, samples containing only observations with complete records are going to be almost surely very small. This implies loss of valuable information, and leads to less efficient estimates.

As a result of the above, it was decided that SHARE would proceed with imputing the missing values of a number of variables in the survey, and this paper discusses the imputation procedures that we have implemented for Release 2.4 of the data for waves 1 and 2 (publicly

available since March 2011).[1] While the vast majority of these procedures were also used in previous joint releases of wave 1 and wave 2 data (i.e., Release 2.3, made available in November 2009, and Release 2.3.1, made available in June 2010), this paper describes the latest modifications that we have made to these procedures for Release 2.4.[2]

Section II of the paper gives details on the prevalence of missing values in SHARE. Section III describes the imputation methodology we have used, while Section IV gives details on implementation issues that are particular to SHARE. Section V concludes.

II. Prevalence of missing values

The first wave of SHARE was conducted in 2004-2005 in eleven countries (Sweden, Denmark, Germany, the Netherlands, Belgium, France, Switzerland, Austria, Italy, Spain, and Greece), while the second wave took place in 2006-2007 and it included, in addition to the aforementioned eleven countries, the Czech Republic, Poland, and Ireland. Imputations are performed for all these countries with the exception of Ireland.[3]

SHARE is a survey that has several different sections recording information on demographics, physical and mental health, cognition, social activities, expectations, employment status and incomes, housing, assets, health expenses, and financial transfers.[4] The sample in each country is representative of the population aged fifty and above, and the second wave contains both a panel and a refresher subsample.

Currently, the imputation procedures in SHARE include a subset of the demographic and economic variables that are recorded in the questionnaire, namely 69 variables in wave 1 and 75 variables in wave 2. In addition, there are a number of economic variables generated during the imputation process that aim to capture magnitudes that are important for the study of numerous topics in both social and biomedical sciences. These variables include, among other things, household income, real and financial assets, and net worth. A complete list of all variables included in the imputation can be found in Appendix Tables A.1 and A.2 for waves 1 and 2, respectively.

[1] The data without imputations are also freely available to the research community from the SHARE website (www.share-project.org).
[2] An earlier description of the SHARE imputation methodology can be found in Christelis (2008).
[3] Israel has also run a survey using the SHARE questionnaire in 2005-2006, and has recently finished collecting the data for a second wave as well. Some simple imputations have already been performed for the first wave for this country, and we plan to implement our full imputation procedure for both waves in the near future.
[4] For more detailed information of SHARE the reader can consult the various chapters in Börsch-Supan, Brugiavini, Jürges, Mackenbach, Siegriest, and Weber (2005), Börsch-Supan and Jürges (2006), and Börsch-Supan, Brugiavini, Jürges, Kapteyn, Mackenbach, Siegriest, and Weber (2008).

The variables included in the imputation process can be further divided into those that are asked at the individual level and those asked at the household level, i.e., to only one person in the household. Among demographic variables, examples of individual-level variables are the level of education, self-reported health status, and the score in a numeracy test, while household-level ones include the location of the house and the number of children and grandchildren. Among economic variables, individual-level variables include earnings from dependent work or self-employment and pension items, while household-level variables include the value of the main residence and the value of food consumed at home.

There are also some variables that can be asked at the individual level to some households, and at the household level to some others. These include most financial assets and financial transfers in wave 1, and their designation as individual- or household-level variables depends on whether the two partners forming the main couple in the household declare to have joint finances or not. In the former case, questions about these items are asked only to the financial respondent, while in the latter case both partners are asked. In wave 2 the question about joint finances is not asked anymore; one partner in the couple is designated as the financial respondent and answers all questions on assets and financial transfers.

The prevalence of missing values in demographic variables can be seen in Tables 1a and 1b for waves 1 and 2, respectively. Information for individual-level variables can be found in columns 1-8 of Table 1a and columns 1-9 of Table 1b. We note that for individual-level demographic variables the prevalence of missing values is typically below 1% of the sample, whereas missing values for household-level demographic variables represent typically less than 3% of the sample (with the exception of the number of grandchildren).

The problem of missing values in individual-level variables is made worse by the fact that in quite a few couples we do not get a response from one of the two partners, not even through a proxy interview.[5] For reasons that will be more extensively discussed in Section IV.2 we have decided to include non-responding partners (NRPs) in our imputation sample. Obviously, this decision increases the prevalence of missing values of individual-level variables.

As NRPs reflect unit non-response, rather than traditional item non-response, we show separately their effects on the prevalence of missing values for individual-level demographic variables in Tables 2a and 2b, which refer to waves 1 and 2, respectively. We note that, with NRPs included, missing values range from 10% to 12% of the sample on

[5] Household-level variables are not affected by this problem, as for them there is one respondent per household.

average, with the problem being more serious in countries with a relatively high percentage of NRPs (e.g. Spain in wave 1).

When assessing the prevalence of missing values for economic variables one needs to take into account the fact that there are typically two decisions that are involved when reporting an amount of an economic variable. The first decision is whether respondents have positive participation (for example if they earn a particular income item or own a particular asset). Subsequently, and conditional on positive participation, we need to determine the value of the corresponding amount. In most cases, the participation question is asked separately from the one referring to the amount, and hence we often have non-missing participation information but missing amount information.

The second issue to keep in mind when considering missingness in economic amounts is related to the nature of the imputation procedure. While the whole sample is relevant for imputing participation, only the sample of participants should be used to impute amounts conditional on participation (non-participants have amounts that are equal to zero). Therefore, one alternative measure of missingness for economic amounts is the ratio of the number of observations with missing values to the number of observations with both missing and non-missing values, conditional on positive participation. As this measure omits the observations of non-participants, and as the values of such observations are overwhelmingly non-missing, ones gets a quite larger prevalence of missing values from this measure than the one obtained from the measure of missingness that is calculated using the whole sample.

However, even if respondents do not give a complete answer to the question about the amount of a particular economic variable, there is still a way to elicit significant information about this value. This is achieved through the mechanism of unfolding brackets: for each economic variable (with the exception of expenditure items), when respondents do not give a complete numerical answer to the amount question, they are subsequently directed to one of three different threshold values (the selection among the three is done randomly). Respondents are then asked if the true value is about equal, higher or lower than the said threshold value. If they report that it is about equal, then their answer is considered complete. If they report that the true value is lower than the threshold value, then they are asked if it is higher, about equal, or lower than the next lower threshold value, and analogously if they report that the true value is higher than the initial threshold value. If this initial value is the lowest of the possible three, and if they report that the true value is lower than that threshold, then no further bracket questions are asked. Once more, a corresponding process exists if the first threshold is the highest one of the three. The three threshold values define four possible

ranges of values, and if respondents finish the bracket process the value of the particular item for which they have positive participation/ownership can be placed in one of the four ranges. This reduces considerably the uncertainly affecting our imputation procedures. Even if respondents do not finish the bracket process (e.g. if they stop after being asked about the first threshold value), they can still give information that excludes from consideration one or more of the four possible ranges of values.

Having all the above in mind, we can now turn to some examples of the prevalence of missing values of economic variables. Specifically, we show results in Table 3a (for wave 1) and Table 3b (for wave 2) for five items: earnings from dependent labor, the main pension, the main residence, bank accounts, and expenditure on food at home. The first two items are individual-level variables in both waves, the value of the main residence and expenditure on food are household-level variables, while the value of bank accounts can be both an individual- and a household-level variable as already described.[6]

The prevalence of missing values, both as a percentage of the total sample (column 1 in both Tables 3a and 3b), and as a percentage of the sample of participants (column 3), depends positively on the likelihood of participation. For example, the high prevalence of home and bank account ownership tends to push the percentage of missing values higher for these two variables. Furthermore, as already mentioned, individual-level variables (like the earnings from dependent labor and the main public pension) tend to have a higher prevalence of missing values than household-level ones. In addition, if the information asked can be possibly considered sensitive (as in the case of bank accounts), then respondents have another motive to not report the value of the amount. On the other hand, given that SHARE respondents who work or receive a public pension are typically fewer than those who own a home, the associated prevalence of missing values for these two income items tends to be smaller, other things being equal.

As a result of the above, bank accounts in wave 1 exhibit the largest percentage of missing values (on average between 35-40% of the total sample, and 40-45% of participants). On the opposite end, the value of the main public pension suffers least from the problem of missing values, which correspond to roughly 5% of the overall sample, and to 10-15% of the sample of participants.

Missing participation (shown in column 2 in both Tables 3a and 3b) is about 0.8% on average for both waves for the case of income from dependent labor, and about 0.4% for the

[6] In wave 2, there are very few cases in which both partners in a couple give complete and differing answers about the value of the bank account. In those cases, the variable is considered an individual-level one.

main public pension. Household-level variables typically have missing participation equal to 2% or less. As bank accounts are often asked at the individual-level in wave 1, the prevalence of their missing values is much higher than in wave 2, in which they are overwhelmingly asked at the household level. Finally, it is assumed that all households spend at least a small amount to buy food-related items, and hence participation for food consumption at home is always assumed to be positive, which also makes it non-missing by definition.

As we have already mentioned, the unfolding brackets procedure mitigates the seriousness of the problem of missing values. We observe that for the household-level variables for which this procedure is implemented (i.e., with the exception of expenditure on food at home), roughly 35% of participants on average finish the bracket sequence (as shown in column 4 of Tables 3a and 3b); hence, the associated variable values can be placed in one of the four possible ranges. The percentage of participants who provide only partial bracket information is relatively small, typically 5-6% or less in both waves.

As expected, including the NRPs in our calculations worsens the problem of missing values in all dimensions (results for individual-level economic variables are shown in Tables 4a and 4b for waves 1 and 2, respectively). The prevalence of missing values for the variables denoting income from dependent labor and from the main public pension rises from about 5% without NRPs to 12-13% on average, while for bank accounts in wave 1 it is between 40-45%. As NRPs do not provide any bracket information by definition, the percentage of respondents who have finished the bracket sequence is lower as well (roughly 20-25% on average).

III. Methodology

The first decision that we had to make about the imputation procedure was whether to use single or multiple imputation (Rubin, 1987). We chose the latter option because imputing a single value for each missing one would result in a complete dataset that would surely be treated by many users in the same way as a dataset with no imputed values whatsoever. As a result, the uncertainty due to the imputation of missing values would not be captured by the estimates generated from the single complete dataset, thus leading to potentially severely underestimated standard errors.

Choosing a multiple imputation procedure also makes it clear that our aim is not to get the best point prediction of the missing value but rather trace the distribution of the possible values, conditional on all the sample information that we can use.

The next decision to be made was how many different implicate datasets to generate, and we decided to generate five, following Rubin's (1987) advice that 3-10 implicates are generally enough for the patterns of missingness typically found in survey data. Five implicates are also the precedent set by the US Survey of Consumer Finances (Kennickell, 1991). The imputation programs are run separately in each of the five implicate datasets; in other words, these datasets are generated independently from one another.

The imputation methodology that we use is the fully conditional specification method (FCS) of van Buuren, Brand, Groothuis-Oudshoorn, and Rubin (2006, henceforth BBGR), and the exposition from this point on follows closely theirs. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_K)$ be a n×K matrix of K variables (all potentially containing missing values) in a sample of size n. $\boldsymbol{Y}$ has a multivariate distribution characterized by a parameter vector $\boldsymbol{\theta}$, denoted by $P(\boldsymbol{Y}; \boldsymbol{\theta})$. The objective of the imputation procedure is to generate imputed values for the missing part of $\boldsymbol{Y}$ (denoted by $\boldsymbol{Y_{mis}}$) that, combined with the non-missing part $\boldsymbol{Y_{obs}}$, will reconstitute as closely as possible the joint distribution $P(\boldsymbol{Y}; \boldsymbol{\theta})$.

One way to proceed would be to assume a fully parametric multivariate density for $\boldsymbol{Y}$, and starting with some priors about $\boldsymbol{\theta}$ to generate imputations of $\boldsymbol{Y_{mis}}$ conditional on $\boldsymbol{Y_{obs}}$ (and on any other vector of variables $\boldsymbol{X}$ that are never missing[7]).

An alternative to specifying a joint multivariate density is to predict any given variable in $\boldsymbol{Y}$, say Y_k, conditional on all remaining variables in the system (denoted by $\boldsymbol{Y_{-k}}$) and a parameter vector $\boldsymbol{\theta_k}$. We apply this procedure to all K variables in $\boldsymbol{Y}$ in a sequential manner, and after the last variable in the sequence has been imputed then a single iteration of this process is considered to be completed. This way the K-dimensional problem of restoring the joint density of $\boldsymbol{Y}$ is broken into K one-dimensional problems of conditional prediction. This breakdown has two principal advantages over the joint approach:

a. It can readily accommodate many different kinds of variables in $\boldsymbol{Y}$ (e.g. binary, categorical, and continuous). This heterogeneity would be very difficult to model with theoretical coherence using a joint distribution of $\boldsymbol{Y}$.
b. It easily allows the imposition of various constraints on each variable (e.g. censoring), as well as constraints across variables. As I will discuss below, both these features are very important in a large household survey like SHARE.

[7] In SHARE the only variables that are essentially never missing are the age and gender of the respondents and the NRPs, as well as the sample stratum to which any observation belongs.

The principal drawback of this method is that there is no guarantee that the K one-dimensional prediction problems lead to convergence to the joint density of $\boldsymbol{Y}$. Because of this potential problem, BBGR ran a number of simulation tests, often complicated by conditions that made imputation difficult, and found that the FCS method performed very well. Importantly, it generated estimates that were generally unbiased, and also good coverage of the nominal confidence intervals.

As the parameter vector $\boldsymbol{\theta}$ of the joint distribution of $\boldsymbol{Y}$ is replaced by the K different parameter vectors $\boldsymbol{\theta_k}$ of the K conditional specifications, BBGR propose to generate the posterior distribution of $\boldsymbol{\theta}$ by using a Gibbs sampler with data augmentation.

Let us suppose that our imputation process has reached iteration *t*, and that we want to impute variable Y_k. We first estimate a statistical model[8] with Y_k as the dependent variable (using only its observed values), and the variables in $\boldsymbol{Y_{-k}}$ as predictors. For every element of $\boldsymbol{Y_{-k}}$ that precedes Y_k in the sequence of variables, its values from iteration *t* are used (i.e., including the imputed ones). On the other hand, for every element of $\boldsymbol{Y_{-k}}$ that follows Y_k in the sequence, its values from iteration *t-1* are used.

After obtaining the parameter vector $\boldsymbol{\theta_k}$ from our estimation, we make a draw $\boldsymbol{\theta_k^*}$ from its posterior distribution[9], i.e., we have

$$\boldsymbol{\theta_k^{*(t)}} \sim P\left(\boldsymbol{\theta_k} | Y_1^{(t)}, \dots, Y_{k-1}^{(t)}, Y_{k,obs}, Y_{k+1}^{(t-1)}, \dots, Y_K^{(t-1)}\right) \qquad (1)$$

The fact that only the observed values of Y_k are used in the estimation constitutes, as BBGR point out, a deviation from most Markov Chain Monte Carlo implementations, and it implies that the estimation sample used for the imputation of any given variable will include only the observations with non-missing values for that variable.

Having obtained the parameter draw $\boldsymbol{\theta_k^{*(t)}}$ at iteration *t* we can use it, together with $\boldsymbol{Y_{-k}^{(t)}}$ and the observed values of Y_k, to make a draw from the conditional distribution of the missing values of Y_k. That is, we have

$$Y_k^{*(t)} \sim P\left(Y_{k,mis} | Y_1^{(t)}, \dots, Y_{k-1}^{(t)}, Y_{k,obs}, Y_{k+1}^{(t-1)}, \dots, Y_K^{(t-1)}; \boldsymbol{\theta_k^{*(t)}}\right) \qquad (2)$$

[8] The model could be a probit, an ordered probit or a linear one, depending on the nature of Y_k.
[9] The formulas used for redrawing the parameter vector can be found in Appendix A of BBGR.

As an example, let us assume that Y_k represents the amount of a particular economic variable, and that we want to impute its missing values at iteration t via ordinary least squares, using the variables in $\boldsymbol{Y}_{-\boldsymbol{k}}^{(t)}$ as predictors. We perform the initial estimation, and obtain the parameter vector $\boldsymbol{\theta}_{\boldsymbol{k}}^{(t)} = (\boldsymbol{\beta}_{\boldsymbol{k}}^{(t)}, \sigma_k^{(t)})$, with $\boldsymbol{\beta}_{\boldsymbol{k}}^{(t)}$ denoting the regression coefficients of $\boldsymbol{Y}_{-\boldsymbol{k}}^{(t)}$, and $\sigma_k^{(t)}$ the standard deviation of the error term. After redrawing the parameter vector $\boldsymbol{\theta}_{\boldsymbol{k}}^{*(t)}$ using (1), we first form a new prediction that is equal to $\boldsymbol{Y}_{-\boldsymbol{k}}^{(t)}\boldsymbol{\beta}_{\boldsymbol{k}}^{*(t)}$. Then, the imputed value $Y_{k,i}^{*(t)}$ for a particular observation i will be equal to $\boldsymbol{Y}_{-\boldsymbol{k},\boldsymbol{i}}^{(t)}\boldsymbol{\beta}_{\boldsymbol{k}}^{*(t)}$ plus a draw of the error term (assumed to be normally distributed with a standard deviation equal to $\sigma_k^{*(t)}$[10]). The error draw for each observation with a missing value for Y_k is made in such a way as to observe any bounds that have been already placed on the admissible values of Y_k for that particular observation. These bounds can have many sources, e.g. they can be the outcomes of the unfolding bracket sequence, overall minima or maxima imposed for the particular variable, or the results of information from another wave.

The process described in (1) and (2) is applied sequentially to all K variables in $\boldsymbol{Y}$, and after the imputation of the last variable in the sequence (i.e., Y_K) iteration t is considered complete. We thus end up with an example of a Gibbs sampler with data augmentation (Tanner and Wong, 1987) that produces the sequence $\{(\boldsymbol{\theta}_{\boldsymbol{1}}^{(t)}, \dots, \boldsymbol{\theta}_{\boldsymbol{K}}^{(t)}, \boldsymbol{Y}_{\boldsymbol{mis}}^{(t)})$: t=1,2,...$\}$. The stationary distribution of this sequence is P($\boldsymbol{Y}_{\boldsymbol{mis}}, \boldsymbol{Y}_{\boldsymbol{obs}}$; $\boldsymbol{\theta}$), provided that convergence of the imputation process is achieved. As Schafer (1997) points out, a sufficient condition for the convergence to the stationary distribution is the convergence of the sequence $\{\boldsymbol{\theta}_{\boldsymbol{1}}^{(t)}, \dots, \boldsymbol{\theta}_{\boldsymbol{K}}^{(t)}\}$ to the conditional distribution of the parameter vector P($\boldsymbol{\theta}|\boldsymbol{Y}_{\boldsymbol{obs}}$), or, equivalently, the convergence of the sequence $\{\boldsymbol{Y}_{\boldsymbol{mis}}^{(t)}\}$ to the conditional distribution of the missing values P($\boldsymbol{Y}_{\boldsymbol{mis}}|\boldsymbol{Y}_{\boldsymbol{obs}}$). Hence, in order to achieve convergence to the stationary distribution of $\boldsymbol{Y}$, we iterate the Gibbs sampler till we have a number of iterations indicating convergence of the distributions of the missing values of all the variables in our system (I discuss further below the methods used for assessing convergence).

One important feature of the FCS method (shared with several other similar approaches found in the imputation literature[11]) is that it operates under the assumption that

[10] In order to make our conditional specifications more compatible with the maintained assumption of normality, the estimation of all models of amounts is done in logarithms.

[11] A similar imputation procedure is proposed by Lepkowski, Raghunathan, Van Hoewyk, and Solenberger (2001). See also BBGR for references to a number of other approaches that have significant similarities to theirs.

the missingness of each variable in $\boldsymbol{Y}$ depends only on other variables in the system and not on the values of the variable itself. This assumption, commonly known as the missing at random (MAR) assumption, is made in the vast majority of imputation procedures applied to large household surveys. It could be argued, however, that it is unlikely to hold for all variables: for example, item non-response in financial assets could depend on whether the respondent owns them in very large or very small values. This would be a case of data missing not at random (MNAR), and, if true, would present major challenges for the construction of the imputation model.

Some evidence on the consequences of the violation of the MAR assumption comes from the results of one of the simulations run by BBGR, which exhibits a NMAR pattern. In addition, BBGR use in this simulation conditional models that are not compatible with a single joint distribution. Even in this rather pathological case, however, the FCS method performs reasonably well, and leads to less biased estimates than an analysis that uses only observations without any missing data. As a result, BBGR conclude that the FCS method (combined with multiple imputation) is a reasonably robust procedure, and that the worry about the incompatibility of the conditional specifications with a joint distribution might be overstated.

One further issue to be addressed is how the iteration process is started, given that, as described above, one needs in any given iteration to use imputed values from the previous iteration. In other words, we need to generate an initial iteration, which will constitute an initial condition that will provide the lagged imputed values to the first iteration. This initial iteration is generated by imputing the first variable in the system based only on variables that are never missing (namely age, gender and geographic location), then the second variable based on the first and the non-missing variables, and so on, till we have a complete set of values for this initial condition. Having obtained this initial set of fully imputed values, we can then start the imputation process using the already described procedures, as denoted in equations (1) and (2).

Once we have obtained the imputed values from the last iteration, we end up with five imputed values for each missing one, i.e., with five different complete datasets that differ from one another only with respect to the imputed values. We then need to consider how to use the five implicate datasets in order to obtain estimates for any magnitude of interest (e.g. descriptive statistics or coefficients of a statistical model).

Let m = 1,…., M index the implicate datasets (with M in our case equal to five) and let $\hat{\beta}_m$ be our estimate of the magnitude of interest from the m[th] implicate dataset. Then the overall estimate derived using all M implicate datasets is just the average of the M separate estimates, i.e.,

$$\overline{\hat{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m \tag{3}$$

The variance of this estimate consists of two parts. Let V_m be the variance of $\hat{\beta}_m$ estimated from the m[th] implicate dataset. Then the within-imputation variance WV is equal to the average of the M variances, i.e.,

$$WV = \frac{1}{M}\sum_{m=1}^{M}V_m \tag{4}$$

One would like each implicate run to explore as much as possible the domain of the joint distribution of the variables in your system; indeed, the possibility of the Markov Chain Monte Carlo process defined in (1) and (2) to jump to any part of this domain is one of the preconditions for its convergence to a joint distribution. This would imply an increased within variance, other things being equal.

The second magnitude one needs to compute is the between-imputation variance BV, which is given by:

$$WV = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}_m - \overline{\hat{\beta}})^2 \tag{5}$$

The between variance is an indicator of the extent to which the different implicate datasets occupy different parts of the domain of the joint distribution of the variables in our system. One would like the implicate runs to not stay far apart but rather mix with one another, thus indicating convergence to the same joint distribution. Therefore, one would like the between variance to be as small as possible relative to the within one.

The total variance TV of our estimate $\bar{\hat{\beta}}$ is equal to:

$$TV = WV + \frac{M+1}{M} BV \tag{6}$$

As Little and Rubin (2002) point out, the second term in (6) indicates the share of the total variance due to missing values. Having computed the total variance, one can perform a t-test of significance using the following formula to compute the degrees of freedom *df* :

$$df = (M-1)\left(1 + \frac{1}{M+1}\frac{WV}{BV}\right)^2 \tag{7}$$

The convergence of our imputation process is the primary factor that determines the number of iterations that our system needs to complete. As already stated, one indication of convergence is the mixing of the five different implicate datasets. Figures 1a and 1b (based on Figure 11.2 in Gelman, Carlin, Stern, and Rubin, 2004) illustrate this point. We have a hypothetical two variable system, consisting of X_1 and X_2 and five implicates. In Figure 1a we have a case in which the five implicates remain very close to their initial values and do not mix at all. Therefore, the between variance is large and the within variance is small (as most of the domain of the joint distribution is not explored). On the other hand, in Figure 1b we have a case in which each implicate moves away from its initial value, and all implicates mix nicely in a space that covers most of the domain of the distribution.

Figures 1a and 1b suggest a couple of possible pitfalls when assessing convergence of the imputation process. First, it is clear that one needs to examine the mixing of the implicates, i.e., whether the between variance is small relative to the within one. Second, looking at how each individual implicate changes over iterations is not a good indicator of convergence: Figure 1a shows that while all five implicates do not change much, there is no convergence of the imputation process. In fact, it is the lack of variability that impedes convergence, as it prevents the five implicates from mixing with one another.

In order to assess the convergence of the imputation process we use the criterion originally proposed by Gelman and Rubin (1992), as restated in Gelman, Carlin, Stern, and Rubin (2004). The criterion can be computed for any magnitude of interest and is equal to

$$GR = \sqrt{\frac{\left((T-1)/T\right)WV + BV}{WV}} = \sqrt{\frac{T-1}{T} + \frac{BV}{WV}} \tag{8}$$

where T is equal to the number of iterations used for its computation. As is clear from (8), the Gelman-Rubin criterion formalizes the intuition that, for convergence to obtain, the between variance has to be small relative to the within one. Gelman, Carlin, Stern, and Rubin (2004) suggest that a value of the criterion below 1.1 is indicative of convergence of the variable in question.

In SHARE, we allow an initial burn-in period, as is the standard practice in the Markov Chain Monte Carlo simulation literature, in order to reduce the dependence of the chain on the initial values. We use five burn-in iterations; hence, we start evaluating the Gelman-Rubin criterion from the seventh iteration on. For each economic variable we typically calculate the criterion for the mean, median and 90th percentile of the distribution of the missing values, and we do the same for a number of composite economic variables as well (e.g. the sum of all pension incomes, and the total value of real and financial assets). In the vast majority of cases we obtain a value of the criterion that indicates convergence pretty early on in the iteration process, namely well before the 15th iteration. In a few cases, however, we have to wait till the 20th iteration or beyond for the value of the criterion to fall sufficiently low. By the 30th iteration all variables in all countries appear to have converged, and hence we stop the imputation process at that point.

An example of quick convergence can be seen in Figure 2, which graphs the Gelman-Rubin criterion for the case of the median value of the main residence of couples in France in wave 1. We see that the critical value of 1.1 is reached by the 11th iteration, and the criterion value falls further in subsequent iterations. The paths of the five different medians are shown in Figure 3; we observe that we have a good mixing of the implicates from very early on in the iteration process.

A case of more difficult convergence is shown in Figure 4 for the value of the main public pension of the partner in couples in Belgium for wave 1. The criterion reaches the critical value at roughly the 20th iteration. From Figure 5, we can see that the five medians mix at the very beginning of the burn-in interval, possibly because the initial condition values were not sufficiently dispersed. Very quickly, however, we observe a deterioration of the mixing, especially for implicate 4, but also for implicate 1. Only in the 12th iteration do we observe a resumption of the mixing of all implicates, and by the 20th iteration this mixing has lasted long enough for the value of the criterion to indicate convergence.

Another way to assess convergence in an informal way is to look at the kernel densities of the imputed values across iterations (for a given implicate). If these distributions

change dramatically in later iterations, this could indicate that convergence to a stable distribution is not yet achieved. As an example, Figure 6 we can see the kernel densities of the imputed values from the third implicate for the expenditure on food at home by couples in Sweden in wave 2. We notice that while the distribution of the missing values in iteration 0 (i.e., the initial condition) is less dispersed than in the remaining iterations, all other densities look reasonably close to one another. We would interpret such stability as possibly a necessary indication for convergence, but not a sufficient one: we always need to assess convergence by looking at the joint evolution of all five implicates.

IV. Implementation issues in SHARE

In the previous Section, the imputation methodology used in SHARE was described in general terms. In this Section, I will discuss some of the particular features of the implementation of this methodology in SHARE.

Before proceeding with the discussion of these features, it is important to point out that imputation in SHARE is done separately for each country. While this choice leads to a reduced number of observations in our estimation samples, it prevents problems that are particular to one country from affecting the imputation in other countries. In addition, it gives us the greatest possible flexibility with respect to the parameters of our estimating equations.

IV.1 Order and Selection of Variables

The Gibbs sampler with data augmentation that was described in Section II involves the prediction of each variable in the system conditional on the remaining ones. Given that this prediction is done sequentially, we need to determine the order with which our variables enter into the Gibbs sampler. As pointed out by Liu, Wong and Kong (1995), this order does not affect the convergence of the Markov chain asymptotically, and the same is true for the frequency with which the prediction of each variable in the sampler is updated. In practice, given that we can allow our imputation model to run for only a relatively limited number of iterations, we need to think carefully whether one choice of variable order over another can improve the convergence of our imputation process. Furthermore, there are practical considerations that impose a particular ordering among some variables.

First, we chose to put the demographic variables before the economic ones in the sequence of variables because the former have typically considerably fewer missing values than the latter. This reduced missingness makes demographic variables good predictors of economic ones in the same iteration.

Second, we put household-level variables after individual-level ones, because in the case of couples we prefer to use the variables of both partners (typically summed up in the case of economic variables) as predictors of household-level variables.

Third, we chose to put some important variables early on in the chain, so as to take advantage of their predictive power for other variables in the same iteration. For example, in the case of demographic variables, we put education and health-related variables early in the sequence, while for the individual-level economic variables we put earnings and the main pension ahead of the remaining ones. For household-level economic variables we gave precedence to the principal residence.

Fourth, there are some logical constraints among variables that dictate their placement in the variable sequence. As we have already mentioned, in the case of economic variables we first determine participation/ownership and then the amount. There are, however, numerous more instances in which we impose logical constraints (a complete list of the constraints is provided in Appendix A.1). For example, we put the missing value of the rent payment equal to zero for home owners. Hence, the variables that have values that can be determined by a logical constraint are put later in the variable sequence than the variables that constitute the source of the constraint. One should note however, that these constraints are imposed only when the relevant values are missing; in other words, we do not use these constraints to change non-missing values.

In addition, while the Gibbs sampler setup implies in principle that every variable in the system should be predicted using all the remaining variables (either from the current or from the last completed iteration), in practice we are occasionally constrained to include a reduced list of predictors. The first reason for this is the sometimes small number of observations in the estimation sample used for the imputation of the amounts of some economic variables. As described in Section II, once participation/ownership of the economic variable is established, the imputation of the amounts proceeds by using in the estimation sample only the observations of owners/participants with non-missing amounts. It turns out that in some cases (e.g., some minor pension items) these observations are fewer than needed for inclusion of the full list of the remaining variables in the system. Hence, we use as predictors only the most important demographic variables (e.g. age, gender, education, self-reported health and numeracy), or variables that are likely to be very good predictors for the item in question. In addition, we group the economic variables into broad categories (e.g. income from all pensions, financial assets). If the usable observations for prediction are

below ten, then we use simple hot-deck to impute missing values; this happens, however, in only a few cases.

A second reason why we might have to use a reduced number of predictors is the lack of convergence of the estimation process when numerous predictors are used. This happens occasionally with the simple probit models used for some variables (e.g. for depression and for participation/ownership of economic variables), and also with the ordered probit models used for some demographic variables (e.g. reading skills, location of the house). Even though the likelihood function of a probit or an ordered probit should in principle converge without problems, in practice convergence is sometimes problematic due to severe collinearity among some regressors, or to the limited variability of some other regressors. If convergence of the likelihood function is not obtained, then the estimation is automatically repeated using a smaller set of predictors, as described above.

We have also chosen to model asset incomes (i.e., incomes from rent, bank accounts, bonds, stocks and mutual funds) separately from the remaining variables in the system, as there are relatively few respondents who earn these incomes, the amounts of which are typically very small. Hence, after the last iteration of the system is completed, we use the other variables in the system as predictors for the asset income items in a one shot imputation, while always taking into account any bracket constraints that we may observe for these income items.

IV.2 Imputation by household kind

One of the first decisions that needed to be made when setting up the imputation procedures in SHARE was how to treat the different kinds of households that can be found in the SHARE sample. The principal differentiating factor between them is whether there is a couple or whether the household head is single (in both cases, there can be more eligible persons in the household, whom we call third respondents).

Due to the problem of NRPs, we treat households headed by couples differently from those headed by singles. The prevalence of NRPs can be seen in Table 5. In wave 1, NRPs range from roughly 5% of the sample in France to 22% in Spain, while in wave 2 the range is between 7% in Greece to 17% in Sweden. Therefore, the problem of NRPs is not negligible in either wave, although it is reduced in wave 2 compared to wave 1, partly because of the incentives given to survey agencies for completing the interviews of both partners in a couple.

One way to deal with the problem of NRPs would be to ignore them, and thus keep them out of the imputation process. A serious problem with this solution comes from the fact that NRPs are unlikely to be missing at random. For example, the second partner in a couple might not respond because (s)he is working and thus has little time to sit down for an interview, or (s)he might be facing health problems that might make an interview difficult. Hence, omitting NRPs that were not missing at random could result in non-representative samples and biased statistical inferences.

A second problem with omitting NRPs altogether is the fact that that income questions in SHARE are asked at the individual level (with the exception of asset incomes), i.e., respondents are not asked to report anything about their partner's income. This has several advantages:

a. responses tend to be more accurate when they reflect only one's personal income situation.
b. individual-level income items can be linked to the respondents' working histories.
c. individual pension incomes can be linked with institutional information taken from SHARE as well as other sources, which makes it easier to draw conclusions about the features of each country's pension system.

The downside of asking income questions at the individual level is that, if one partner in the couple does not respond, then it becomes difficult to get an accurate measure of total household income, which is a very important piece of information that, as already mentioned, is needed for the study of numerous issues in social and biomedical sciences.

As a result of the aforementioned concerns, it was decided that NRPs were going to be included in the SHARE imputation sample. We tried however, to reduce the need to impute information about NRPs in a number of ways. First, we used information on NRPs from another wave: 1,202 NRPs in wave 1 (31% of all wave 1 NRPs) are interviewed in wave 2, while 1,127 NRPs in wave 2 (26% of all wave 2 NRPs) are interviewed in wave 1.[12] As I will discuss in more detail in Section IV.3, a full interview in a different wave can provide a lot of information about NRPs. Second, we asked in wave 2 some questions at the household level, namely on assets and on financial transfers, irrespective of whether the couple had separate or joint finances. We made this choice because it was likely that the household financial respondent knew enough about these items to give an accurate answer for the couple as a whole (these questions are asked at the household level also in other major surveys like the

[12] These NRPs do not include any wave 1 respondents that passed away before the wave 2 interview.

US Survey of Consumer Finances, and the Health and Retirement Study). Third, in wave 2 there were a number of questions about a NRP that were asked to the responding partner, namely questions about years of education, current employment status, and work history. Fourth, in wave 2 there was a question asked about total household income in the month of the interview, which could be used to deduce (some of) the income items of the NRP.

Having decided to include the NRPs in the imputations, we needed to think how to impute their missing information. First, it is important to recall that we have information about the responding partner, which could be used as predictor for the missing information of the NRP. For example, the education level of the responding partner can be informative about that of the NRP due to assortative matching, and similar arguments can hold about cognition, working status, and income levels. As a result, for each variable to be imputed in households with couples, other variables corresponding to both partners are used as predictors. This in turn implies that imputation for couples is done separately from that for singles because for the latter predictors can come only from the respondent (singles do not have a partner). The downside of doing the imputation by household kind is that the samples used in our estimation become smaller.

Having separate imputation processes for couples and singles allows us to simplify the treatment of demographic variables for singles. As there are no NRPs for them, the prevalence of missing data for the demographic variables is very small. Therefore, we use simple hot-deck to impute missing values for those variables, with the conditioning variables tailored to each case, but typically including age, gender and education. Then we use the fully imputed demographic variables as predictors for the economic ones. On the other hand, in the case of couples demographic variables are fully integrated into the Gibbs sampler described in Section II.

Finally, we also decided to treat third respondents separately, given their very limited prevalence: there are only 336 of them in wave 1 (1.04% of the total sample) and 206 in wave 2 (0.55% of the total sample). The imputation for third respondents was performed using simple hot-deck by age, gender and education. As there were a few cases for which third respondents were chosen as the main respondents for specific household-level economic variables, their responses were also used in the imputation process for the main couple in the household or for the single head.

IV.3 Linking observations across waves

Given that we had two waves of data available, we tried to use for the imputation of a given wave as much information as possible from the other wave. This information was needed especially for the case of NRPs. As already described in Section IV.2, in wave 2 we used a number of questions that could be used to fill in missing information for an NRP in wave 1. For example, if in wave 2 a wave 1 NRP reported that she was currently working and that she had started working at that job before the time she was supposed to be interviewed in wave 1, then in wave 1 she is also considered to be working, and thus we impute earnings to her. The same procedure is followed for many pension items, for which we can also use some other logical constraints for deducing participation. For example, if the respondent does not get a particular pension in wave 2, then she is also very unlikely to get it in wave 1, as pensions are almost never discontinued.

While this information is crucial for determining participation, we can also get some information about missing wave 1 amounts from a complete wave 2 answer. For example, if the person has worked in the same job in both waves and we know her salary in wave 2, then we can reasonably infer that her wave 1 salary is equal to the wave 2 one plus or minus a given percentage. This percentage is calculated, for a given country, from the observations that have complete information in both waves. We use this calculated interval for the wave 1 salary together with any other available information about the allowed range of values (e.g. from brackets, or any institutional minima or maxima), so as to tighten the final allowed range of values for the wave 1 salary.

Obviously, we can use similar procedures also going forward in time, i.e., from wave 1 to wave 2. For example, for some pension items we can impose logical constraints on participation going forward in time: if a respondent gets a pension in wave 1, then she almost surely gets it in wave 2 as well.

In addition to getting participation and amount information from combining waves, we also had to consider how to use this information in our estimation. The first possibility was to do a two-wave panel estimation for the items that were common across waves. This would allow us to get larger estimation samples and thus use more information in our prediction. The second possibility was to do a cross-sectional estimation for wave 1, and then use for each variable in wave 2 its lagged value from wave 1 as an additional predictor. This increases significantly the predictive accuracy of our equations given the large persistence typically observed in both demographic and economic variables. Obviously, as we had only two waves at our disposal, we could not use the lagged dependent variable in a full-blown

panel estimation. The downside of using the lagged dependent variable as a predictor is the smaller size of our estimation samples compared to the one we could obtain if we performed a two-wave panel estimation. In both cases, we have to do a separate estimation for the panel and the refresher sample (which is typically quite smaller than the panel one). In the end we opted for the increased predictive power of the lagged dependent variable.

Given that SHARE is an ongoing survey, one could in principle combine both methods using the third and subsequent waves, i.e., one could perform a panel imputation procedure using a lagged dependent variable. It is very difficult to use such an approach, however, because the third wave of SHARE (SHARELIFE) is a retrospective survey that is fundamentally different in its questionnaire from the first two waves; hence, it cannot be easily integrated into the existing imputation process. From the fourth wave on (scheduled to go into the field in early 2011), the questionnaire reverts more or less to its old format. Therefore, one could conceivably use the second wave variables as lagged dependent variables in the fourth wave, which would imply a two-wave time distance instead of the one-wave time distance currently present between the second and first waves.

All in all, because of the discontinuity in the questionnaire design, we think that it is probably more practical to do a cross-sectional estimation in each wave using a lagged dependent variable when possible, rather than attempt a full-blown panel estimation.

IV.4 Problems affecting earnings from dependent labor

An important variable in our imputation system, namely earnings from dependent employment in the year prior to the interview, is affected by two problems. The first problem is that for some respondents the value of the amount was set to zero even though they indicated that they were working. The prevalence of this problem can be seen in columns 1-2 and 5-6 of Table 6 for waves 1 and 2, respectively. While for wave 1 the problem is not really widespread for any country, its prevalence in wave 2 is non-negligible in Sweden, Belgium, Switzerland, Italy, and Greece.

One possible reason for this problem could be that before the question about the earnings from last year was asked, there was another question asking about the amount of the last payment received prior to the interview. Hence, we conjecture that at least some respondents were confused and thought that the second question (about earnings in the previous year) referred to any earnings that were additional to those that were asked about in the first question. Given that the vast majority of respondents has only one source of earnings

from dependent labor, this confusion could have led some of them to report zero amounts in the second question.

The second problem affecting the variable denoting last year's earnings is that some respondents reported very similar numbers to both earnings questions.[13] Once more, they might have been confused and thus thought that the second question asked about the same concept of earnings as the first one. The prevalence of this second problem is shown in columns 3-4 and 5-6 of Table 6 for waves 1 and 2, respectively. We can see that it affects most countries in the sample, and is especially pronounced in Switzerland.

While the first problem was corrected from the first joint release of the first two waves (Release 2.3), the second problem was not corrected till Release 2.4. As a result, in a number of countries the distribution of earnings before Release 2.4 had a double peak, with the first peak being at low values of income, as the last payment (typically the monthly income) was reported instead of the yearly income. This pattern can be seen clearly in Figures 7 and 8 for waves 1 and 2, respectively.

In the case of respondents that did not change jobs between the year prior to the interview and the time of the interview, the problem was corrected by using the reported value of the last payment prior to the interview and annualizing it for the previous year, after allowing for additional payments and related bonuses.[14] This correction was applied outside the imputation process, as we think that it will result in less noisy estimates than those obtained from a full imputation that did not take into account the amount of this payment. The results of this correction can be seen again in Figures 7 and 8, where the double peaks once present in many countries (notably Germany, Belgium, Switzerland, Austria, Italy and Spain in wave 1, and Germany, the Netherlands, Belgium, Switzerland, Italy and the Czech Republic in wave 2) are much less prominent in Release 2.4 data.

Another way to look at the effects of this correction is to examine what happens at the low quantiles of the distribution of earnings from dependent labor (conditional on participation), data for which are shown in Table 7. As expected, the bottom quantiles are much more affected by the correction than the median or the 75th quantile. In other words, while there is a general movement of the frequency distribution to the right, this movement is much more pronounced for the bottom quantiles.

We also examined how the correction affected the imputation of other economic variables for the household, namely total income, the value of the home, food consumption

[13] We are grateful to Thomas Georgiadis for alerting us to this issue.
[14] Omar Paccagnella kindly provided these calculations.

and net worth. We could detect only a small effect on total household income (most notably in Switzerland in both waves), probably because respondents who are still working are a minority in our sample, which consists of those aged fifty and above. As for the other economic variables, we did not notice any significant changes between the two data releases that could be attributed to this earnings correction.

V. Conclusion

Like all major household surveys, SHARE suffers from item non-response. In this paper, we have described the procedures that we have used to impute the resulting missing values. We have performed our imputation using an iterative conditional specification approach that has been used, with some variation, in many other household surveys. We have also paid special attention to the issue of convergence of our imputation process, and to that effect we have used the Gelman-Rubin convergence criterion, together with other less formal approaches (e.g. inspection of kernel density functions across iterations).

Given that SHARE is a multi-country survey that has many different questionnaire sections, it presents us with several complications that necessitate some adjustments to the imputation framework of BBGR, especially with respect to the selection of the variables used as predictors in our estimating equations. Overall, however, we have tried to keep departures from the BBGR framework to a minimum.

In the future, we will attempt to make more extended use of information from future survey waves during the imputation procedure of a given wave, even in a cross-sectional imputation setting. For example, instead of using only the lagged dependent variable as a predictor in our estimation, we will try to find ways to use one or more of its future values as predictors as well.

Ultimately, however, the best way to deal with the problem of missing values is to reduce their prevalence, and thus the need for any imputation. In SHARE, the most important step in this direction would be the reduction of the number of NRPs. While progress has been made on that front in wave 2, we are still trying different approaches that will hopefully further reduce the extent of the problem. In addition, given that SHARE has a large panel component, we are considering new ways to use information from different waves (especially the life history information from SHARELIFE), in order to reduce the uncertainty affecting our imputations.

References

Börsch-Supan, A., A. Brugiavini, H. Jürges, J. Mackenbach, J. Siegriest and G. Weber, eds. (2005). *The Survey on Health, Aging and Retirement in Europe*. Mannheim: Mannheim Research Institute for the Economics of Aging.

Börsch-Supan, A. and H. Jürges, eds. (2006). *The Survey on Health, Aging and Retirement in Europe-Methodology*. Mannheim: Mannheim Research Institute for the Economics of Aging.

Börsch-Supan, A., A. Brugiavini, H. Jürges, A. Kapteyn, J. Mackenbach, J. Siegriest, and G. Weber (eds). (2008). *First Results from the Survey on Health, Aging and Retirement in Europe (2004-2007) : Starting the Longitudinal Dimension*. Mannheim: Mannheim Research Institute for the Economics of Aging.

van Buuren, S., J.P.L. Brand, C.G.M. Groothuis-Oudshoorn, and D.B. Rubin. (2006). "Fully Conditional Specification in Multivariate Imputation." *Journal of Statistical Computation and Simulation*, 76:1049-64.

Christelis, D. (2008). "Item Non-response in SHARE Wave 2". In *First Results from the Survey on Health, Aging and Retirement in Europe (2004-2007) : Starting the Longitudinal Dimension*. A. Börsch-Supan, A. Brugiavini, H. Jürges, A. Kapteyn, J. Mackenbach, J. Siegriest and G. Weber, eds. Mannheim: Mannheim Research Institute for the Economics of Aging.

Gelman, A., and D.B. Rubin. (1992). "Inference from Iterative Simulation using Multiple Sequences." *Statistical Science*, 7: 457-511.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. (2004). *Bayesian Data Analysis*, *Second Edition*. Boca Raton, FL: Chapman and Hall.

Kennickell, A.B. (1991). "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation." *Proceedings of the Section on Survey Research Methods*, 1991 Annual Meeting of the American Statistical Association, Atlanta, GA.

Lepkowski, J. M., T. E. Raghunathan, J. Van Hoewyk, and P. Solenberger. (2001). "A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models." *Survey Methodology*, 27: 85-95.

Little, R. E. and D. B. Rubin. (2002). *Statistical Analysis of Missing Data,* 2nd Edition. New York, NY: John Wiley & Sons.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman and Hall.

Tanner, M.A. and W.H. Wong. (1987). "The Calculation of Posterior Distributions by Data Augmentation (with discussion). " *Journal of the American Statistical Association*, 82: 528-550.

Table 1a. Prevalence of missing values in demographic variables in wave 1, excluding NRPs

Country	(1) Education	(2) Self-Reported Health	(3) Limited in Usual Activities due to Health	(4) Number of Limitations in Activities of Daily Living	(5) Number of Limitations in Instrumental Activities of Daily Living	(6) Felt Depressed in the Previous Month	(7) Numeracy Score	(8) Self-Assessed Reading Skills	(9) Number of Rooms in the House	(10) Location of the House	(11) Family Makes Ends Meet	(12) Number of Children	(13) Number of Grandchildren
Sweden	0.13	0.23	0.16	0.20	0.20	1.18	0.69	1.38	1.12	0.75	1.59	0.00	4.67
Denmark	0.18	0.23	0.23	0.29	0.29	0.88	0.76	1.23	1.28	1.28	1.96	0.31	6.66
Germany	0.07	0.17	0.17	0.20	0.20	1.20	0.73	1.26	2.30	1.80	2.35	0.36	10.78
Netherlands	0.47	0.64	0.57	0.67	0.67	2.18	1.01	2.75	2.00	1.69	2.20	0.36	7.33
Belgium	0.10	0.21	0.24	0.29	0.29	0.50	0.50	0.52	0.99	0.59	1.26	0.12	8.55
France	1.60	2.51	3.01	2.60	2.60	5.54	4.01	5.39	2.89	2.37	3.27	1.14	8.09
Switzerland	0.10	0.40	0.40	0.40	0.40	0.50	0.80	0.50	3.09	1.69	3.65	0.70	10.80
Austria	0.16	0.32	0.32	0.42	0.42	0.79	0.48	0.95	0.64	0.43	0.85	0.56	9.42
Italy	0.16	0.35	0.31	0.31	0.31	0.43	0.51	0.66	1.57	1.01	1.91	0.47	6.02
Spain	0.04	0.63	0.54	0.71	0.71	1.92	1.13	2.13	4.56	3.02	4.51	0.66	7.07
Greece	0.03	0.10	0.14	0.10	0.10	2.48	0.24	2.76	0.30	0.15	0.55	0.00	6.12

Notes: All values are expressed in percentages.

Table 1b. Prevalence of missing values in demographic variables in wave 2, excluding NRPs

Country	(1) Education	(2) Self-Reported Health	(3) Risk Preferences	(4) Limited in Usual Activities due to Health	(5) Number of Limitations in Activities of Daily Living	(6) Number of Limitations in Instrumental Activities of Daily Living	(7) Felt Depressed in the Previous Month	(8) Numeracy Score	(9) Self-Assessed Reading Skills	(10) Number of Rooms in the House	(11) Location of the House	(12) Family Makes Ends Meet	(13) Number of Children	(14) Number of Grandchildren
Sweden	0.87	0.15	3.55	0.07	0.15	0.15	2.19	1.24	4.93	2.68	3.96	4.16	0.67	5.47
Denmark	0.73	0.23	3.18	0.19	0.27	0.27	1.11	0.84	1.16	0.82	3.52	3.64	0.09	5.66
Germany	1.25	0.23	3.22	0.23	0.23	0.23	1.05	1.29	4.24	3.10	3.89	4.00	0.00	9.91
Netherlands	1.20	0.56	2.76	0.56	0.56	0.56	1.43	0.90	5.35	3.14	4.44	4.50	0.23	7.34
Belgium	0.79	0.03	2.88	0.03	0.06	0.06	1.10	0.35	4.28	1.71	2.76	2.90	0.15	8.64
France	3.74	1.52	5.44	1.62	1.58	1.58	4.21	3.81	4.98	3.18	4.08	4.13	0.65	7.93
Switzerland	0.96	0.48	2.04	0.55	0.41	0.41	0.82	0.75	1.42	1.87	2.42	2.42	0.00	10.69
Austria	1.57	0.07	1.50	0.07	0.15	0.15	0.37	0.37	20.00	3.06	2.13	2.33	0.00	7.86
Italy	1.14	0.20	3.21	0.20	0.20	0.20	0.64	0.54	1.20	1.59	1.59	1.75	0.14	6.58
Spain	4.40	0.13	11.97	0.09	0.13	0.13	2.42	1.30	4.19	4.76	5.16	6.29	0.10	6.79
Greece	4.59	0.40	3.48	0.12	0.34	0.34	1.30	0.40	1.97	1.61	1.24	1.43	0.23	6.42
Czech Republic	1.63	0.25	3.91	0.25	0.32	0.32	1.70	0.39	1.70	1.04	1.90	1.70	0.00	4.41
Poland	2.47	0.41	3.67	0.36	0.49	0.49	1.42	0.81	1.38	1.30	1.47	1.69	0.77	3.67

Notes: All values are expressed in percentages.

Table 2a. Prevalence of missing values in demographic variables in wave 1, including NRPs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Country	**Education**	**Self-Reported Health**	**Limited in Usual Activities due to Health**	**Number of Limitations in Activities of Daily Living**	**Number of Limitations in Instrumental Activities of Daily Living**	**Felt Depressed in the Previous Month**	**Numeracy Score**	**Self-Assessed Reading Skills**
Sweden	10.44	15.46	15.40	15.43	15.43	16.26	15.85	16.43
Denmark	4.32	6.79	6.79	6.84	6.84	7.39	7.28	7.72
Germany	9.65	12.70	12.70	12.73	12.73	13.60	13.20	13.66
Netherlands	9.89	12.09	12.03	12.12	12.12	13.45	12.41	13.96
Belgium	6.91	8.94	8.97	9.01	9.01	9.20	9.20	9.23
France	5.84	7.71	8.18	7.80	7.80	10.58	9.13	10.44
Switzerland	9.00	12.59	12.59	12.59	12.59	12.67	12.94	12.67
Austria	8.21	11.99	11.99	12.08	12.08	12.41	12.13	12.55
Italy	11.96	18.27	18.24	18.24	18.24	18.33	18.40	18.53
Spain	13.57	22.34	22.28	22.41	22.41	23.35	22.73	23.52
Greece	6.91	7.42	7.45	7.42	7.42	9.63	7.55	9.88

Notes: All values are expressed in percentages.

Table 2b. Prevalence of missing values in demographic variables in wave 2, including NRPs

Country	(1) Education	(2) Self-Reported Health	(3) Risk Preferences	(4) Limited in Usual Activities due to Health	(5) Number of Limitations in Activities of Daily Living	(6) Number of Limitations in Instrumental Activities of Daily Living	(7) Felt Depressed in the Previous Month	(8) Numeracy Score	(9) Self-Assessed Reading Skills
Sweden	2.75	17.12	3.55	17.05	17.12	17.12	18.81	18.02	35.42
Denmark	1.79	8.61	3.18	8.58	8.65	8.65	9.42	9.17	13.07
Germany	2.98	12.23	3.22	12.23	12.23	12.23	12.95	13.16	22.84
Netherlands	4.73	16.48	2.76	16.48	16.48	16.48	17.20	16.76	29.01
Belgium	1.03	9.41	2.88	9.41	9.44	9.44	10.38	9.69	38.01
France	4.55	11.43	5.44	11.52	11.49	11.49	13.85	13.49	21.91
Switzerland	6.46	16.14	2.04	16.20	16.08	16.08	16.43	16.37	23.59
Austria	1.92	11.14	1.50	11.14	11.21	11.21	11.41	11.41	57.98
Italy	1.42	10.17	3.21	10.17	10.17	10.17	10.56	10.47	18.77
Spain	4.66	10.53	11.97	10.49	10.53	10.53	12.59	11.58	16.76
Greece	4.61	7.45	3.48	7.19	7.39	7.39	8.28	7.45	17.05
Czech Republic	5.73	7.59	3.91	7.59	7.66	7.66	8.94	7.73	8.94
Poland	4.52	16.54	3.67	16.51	16.61	16.61	17.39	16.88	17.36

Notes: All values are expressed in percentages.

Table 3a. Missing values in economic variables in wave 1, excluding NRPs

Country	(1) Missing Values (% of the Total Sample)	(2) Missing Participation /Ownership (% of the Total Sample)	(3) Missing Amounts (% of Owners / Participants)	(4) Full Bracket Information (% of Observations with Missing Amounts)	(5) Partial Bracket Information (% of Observations with Missing Amounts)	(6) No Bracket Information (% of Observations with Missing Amounts)
Panel A. Income from Dependent Labor						
Sweden	1.64	0.26	3.06	25.58	2.33	72.09
Denmark	2.69	0.53	5.06	19.05	2.38	78.57
Germany	9.84	0.66	24.05	24.29	2.86	72.86
Netherlands	5.67	0.87	13.47	30.82	2.74	66.44
Belgium	5.93	0.52	18.90	36.15	2.82	61.03
France	7.45	3.32	13.88	31.01	1.27	67.72
Switzerland	7.67	1.29	17.56	33.33	4.35	62.32
Austria	4.49	0.69	16.41	13.51	6.76	79.73
Italy	2.97	0.55	14.82	29.85	1.49	68.66
Spain	6.01	0.92	20.29	42.52	3.94	53.54
Greece	7.07	0.76	20.19	21.93	5.88	72.19
Panel B. Main Public Pension Income						
Sweden	4.26	0.16	9.83	48.44	4.69	46.88
Denmark	4.80	0.41	13.41	49.37	2.53	48.10
Germany	8.88	0.70	17.92	36.11	7.94	55.95
Netherlands	1.24	0.57	2.87	9.68	0.00	90.32
Belgium	7.92	0.24	20.41	47.14	3.03	49.83
France	9.99	1.85	18.88	28.52	2.75	68.73
Switzerland	5.68	0.50	12.05	12.96	0.00	87.04
Austria	8.14	0.48	14.45	31.08	2.70	66.22
Italy	2.85	0.51	11.07	42.62	4.92	52.46
Spain	5.43	0.58	16.74	45.00	0.83	54.17
Greece	4.07	0.21	11.52	27.19	4.39	68.42
Panel C. Main Residence						
Sweden	6.50	0.56	8.75	41.18	1.47	57.35
Denmark	6.46	0.68	9.32	25.33	4.00	70.67
Germany	12.14	1.55	21.96	48.10	8.44	43.46
Netherlands	6.50	1.54	9.40	49.53	3.74	46.73
Belgium	14.18	0.36	17.91	58.22	5.29	36.49
France	21.66	2.23	29.08	49.32	7.99	42.69
Switzerland	13.90	1.83	23.81	51.11	4.44	44.44
Austria	12.21	0.43	21.18	40.83	2.96	56.21
Italy	19.57	1.12	24.61	51.02	3.50	45.48
Spain	25.21	2.22	28.60	52.17	1.60	46.22
Greece	16.04	0.15	19.15	38.99	3.77	57.23

Table 3a (continued). Missing values in economic variables in wave 1, excluding NRPs

Country	(1) Missing Values (% of the Total Sample)	(2) Missing Participation /Ownership (% of the Total Sample)	(3) Missing Amounts (% of Owners / Participants)	(4) Full Bracket Information (% of Observations with Missing Amounts)	(5) Partial Bracket Information (% of Observations with Missing Amounts)	(6) No Bracket Information (% of Observations with Missing Amounts)
			Panel D. Bank Accounts			
Sweden	22.67	1.67	23.42	43.30	2.89	53.81
Denmark	28.98	2.97	33.09	40.83	1.94	57.22
Germany	44.34	7.21	46.92	47.49	2.28	50.23
Netherlands	35.50	3.70	37.71	46.06	3.60	50.33
Belgium	56.16	5.20	59.77	38.70	2.32	58.98
France	46.14	6.78	49.18	53.56	2.34	44.11
Switzerland	40.05	7.73	42.96	45.73	5.12	49.15
Austria	32.84	2.19	43.27	40.29	2.04	57.67
Italy	28.66	3.64	44.65	48.70	2.60	48.70
Spain	46.56	5.08	55.69	55.41	1.32	43.27
Greece	30.99	5.92	48.95	20.34	2.59	77.07
			Panel E. Consumption of Food at Home			
Sweden	7.01	0.00	7.01	-..-	-..-	-..-
Denmark	19.98	0.00	19.98	-..-	-..-	-..-
Germany	12.54	0.00	12.54	-..-	-..-	-..-
Netherlands	13.87	0.00	13.87	-..-	-..-	-..-
Belgium	34.04	0.00	34.04	-..-	-..-	-..-
France	28.82	0.00	28.82	-..-	-..-	-..-
Switzerland	18.82	0.00	18.82	-..-	-..-	-..-
Austria	12.07	0.00	12.07	-..-	-..-	-..-
Italy	17.77	0.00	17.77	-..-	-..-	-..-
Spain	28.58	0.00	28.58	-..-	-..-	-..-
Greece	8.58	0.00	8.58	-..-	-..-	-..-

Table 3b. Missing values in economic variables in wave 2, excluding NRPs

Country	(1) Missing Values (% of the Total Sample)	(2) Missing Participation /Ownership (% of the Total Sample)	(3) Missing Amounts (% of Owners / Participants)	(4) Full Bracket Information (% of Observations with Missing Amounts)	(5) Partial Bracket Information (% of Observations with Missing Amounts)	(6) No Bracket Information (% of Observations with Missing Amounts)
Panel A. Income from Dependent Labor						
Sweden	3.17	0.66	6.24	50.00	2.78	47.22
Denmark	2.98	0.54	5.41	50.00	3.03	46.97
Germany	5.49	0.39	17.43	29.32	3.01	67.67
Netherlands	6.09	0.83	15.84	26.03	4.11	69.86
Belgium	3.22	0.13	13.00	36.73	1.02	62.24
France	5.26	2.33	10.67	49.50	3.96	46.53
Switzerland	7.18	0.62	16.89	34.34	5.05	60.61
Austria	1.79	0.22	10.58	18.18	4.55	77.27
Italy	1.68	0.40	8.01	25.64	5.13	69.23
Spain	6.28	0.67	28.35	30.53	8.40	61.07
Greece	5.64	0.40	22.31	10.47	1.16	88.37
Czech Republic	8.55	0.39	23.83	56.41	2.99	40.60
Poland	3.24	0.73	11.19	41.54	3.08	55.38
Panel B. Main Public Pension Income						
Sweden	2.55	0.18	4.99	37.88	15.15	46.97
Denmark	2.10	0.31	5.09	25.49	0.00	74.51
Germany	6.70	0.12	13.48	34.88	6.40	58.72
Netherlands	3.16	0.49	7.43	24.68	6.49	68.83
Belgium	5.74	0.03	14.27	17.13	3.87	79.01
France	8.02	1.08	14.30	34.26	5.56	60.19
Switzerland	3.76	0.21	8.35	23.08	0.00	76.92
Austria	4.03	0.15	6.87	26.42	7.55	66.04
Italy	1.64	0.23	4.46	20.45	4.55	75.00
Spain	4.58	0.27	13.71	22.45	8.16	69.39
Greece	5.15	0.22	14.94	26.83	3.05	70.12
Czech Republic	8.83	0.21	15.56	50.20	3.64	46.15
Poland	3.85	0.45	6.65	5.62	0.00	94.38
Panel C. Main Residence						
Sweden	7.41	1.62	9.62	34.75	2.84	62.41
Denmark	3.75	0.63	4.82	29.51	8.20	62.30
Germany	13.37	2.53	19.58	39.02	3.90	57.07
Netherlands	8.50	3.09	11.46	16.30	5.93	77.78
Belgium	11.70	1.68	14.15	52.08	5.42	42.50
France	25.58	3.11	33.24	45.89	11.82	42.28
Switzerland	10.78	1.67	17.17	34.58	2.80	62.62
Austria	18.24	1.42	26.66	21.97	4.05	73.99
Italy	16.84	1.54	19.86	51.32	6.91	41.78
Spain	35.50	4.38	39.35	48.49	4.43	47.08
Greece	21.45	1.24	24.30	21.23	4.38	74.40
Czech Republic	16.93	0.46	22.82	54.01	7.10	38.89
Poland	22.47	1.36	30.31	56.51	2.86	40.63

Table 3b (continued). Missing values in economic variables in wave 2, excluding NRPs

Country	(1) Missing Values (% of the Total Sample)	(2) Missing Participation / Ownership (% of the Total Sample)	(3) Missing Amounts (% of Owners / Participants)	(4) Full Bracket Information (% of Observations with Missing Amounts)	(5) Partial Bracket Information (% of Observations with Missing Amounts)	(6) No Bracket Information (% of Observations with Missing Amounts)
			Panel D. Bank Accounts			
Sweden	22.46	3.14	22.47	29.16	5.30	65.54
Denmark	20.72	1.65	21.22	27.61	4.79	67.61
Germany	37.13	3.45	38.02	29.85	8.62	61.53
Netherlands	33.41	3.36	33.45	25.73	6.35	67.92
Belgium	49.14	2.60	50.44	33.65	4.34	62.01
France	47.76	4.53	48.47	47.93	5.90	46.17
Switzerland	34.20	2.22	35.17	33.05	6.72	60.22
Austria	32.70	2.40	36.13	28.84	2.19	68.97
Italy	29.92	2.22	36.57	32.19	6.15	61.66
Spain	46.76	5.41	55.90	29.51	5.05	65.44
Greece	29.51	7.82	55.79	17.96	4.44	77.59
Czech Republic	32.58	2.50	48.99	32.84	7.06	60.10
Poland	14.10	2.09	47.49	22.94	2.75	74.31
			Panel E. Consumption of Food at Home			
Sweden	10.04	0.00	10.04	-..-	-..-	-..-
Denmark	15.85	0.00	15.85	-..-	-..-	-..-
Germany	10.86	0.00	10.86	-..-	-..-	-..-
Netherlands	12.27	0.00	12.27	-..-	-..-	-..-
Belgium	18.76	0.00	18.76	-..-	-..-	-..-
France	23.35	0.00	23.35	-..-	-..-	-..-
Switzerland	11.25	0.00	11.25	-..-	-..-	-..-
Austria	7.45	0.00	7.45	-..-	-..-	-..-
Italy	8.74	0.00	8.74	-..-	-..-	-..-
Spain	19.76	0.00	19.76	-..-	-..-	-..-
Greece	6.08	0.00	6.08	-..-	-..-	-..-
Czech Republic	10.92	0.00	10.92	-..-	-..-	-..-
Poland	13.19	0.00	13.19	-..-	-..-	-..-

Table 4a. Missing values in economic variables in wave 1, including NRPs

Country	(1) Missing Values (% of the Total Sample)	(2) Missing Participation /Ownership (% of the Total Sample)	(3) Missing Amounts (% of Owners / Participants)	(4) Full Bracket Information (% of Observations with Missing Amounts)	(5) Partial Bracket Information (% of Observations with Missing Amounts)	(6) No Bracket Information (% of Observations with Missing Amounts)
Panel A. Income from Dependent Labor						
Sweden	15.63	12.93	13.19	5.31	0.48	94.20
Denmark	8.92	6.08	9.84	9.30	1.16	89.53
Germany	20.70	12.09	31.15	17.00	2.00	81.00
Netherlands	16.04	11.17	21.83	17.18	1.53	81.30
Belgium	13.54	7.92	27.17	22.58	1.76	75.66
France	11.95	7.71	16.81	24.75	1.01	74.24
Switzerland	18.36	11.54	25.35	20.91	2.73	76.36
Austria	15.25	11.19	21.13	9.90	4.95	85.15
Italy	18.59	16.06	31.62	11.24	0.56	88.20
Spain	23.09	17.91	40.17	16.12	1.49	82.39
Greece	13.30	7.42	23.42	18.14	4.87	76.99
Panel B. Main Public Pension Income						
Sweden	11.16	4.97	19.09	22.38	2.17	75.45
Denmark	7.28	2.35	16.39	39.00	2.00	59.00
Germany	15.38	6.57	23.63	25.49	5.60	68.91
Netherlands	7.28	5.26	9.88	2.61	0.00	97.39
Belgium	11.23	3.00	24.17	37.94	2.44	59.62
France	12.45	3.56	21.83	23.78	2.29	73.93
Switzerland	11.19	4.37	17.57	8.33	0.00	91.67
Austria	13.81	4.62	20.15	20.81	1.81	77.38
Italy	12.24	7.79	27.08	14.29	1.65	84.07
Spain	15.13	7.63	33.44	18.00	0.33	81.67
Greece	7.77	3.29	16.41	18.02	2.91	79.07
Panel C. Bank Accounts						
Sweden	27.38	7.66	25.92	37.84	2.52	59.64
Denmark	31.76	6.76	34.12	38.99	1.86	59.15
Germany	47.02	11.66	48.12	45.27	2.18	52.56
Netherlands	40.37	10.98	40.36	41.22	3.23	55.56
Belgium	57.75	8.62	60.16	38.08	2.28	59.64
France	46.81	7.93	49.35	53.18	2.32	44.50
Switzerland	44.09	13.95	43.95	43.93	4.92	51.15
Austria	37.31	8.70	44.74	37.96	1.93	60.12
Italy	36.14	13.69	46.77	44.71	2.39	52.90
Spain	53.11	16.72	58.29	49.84	1.19	48.97
Greece	33.02	8.69	49.50	19.90	2.53	77.57

Table 4b. Missing values in economic variables in wave 2, including NRPs

Country	(1) Missing Values (% of the Total Sample)	(2) Missing Participation /Ownership (% of the Total Sample)	(3) Missing Amounts (% of Owners / Participants)	(4) Full Bracket Information (% of Observations with Missing Amounts)	(5) Partial Bracket Information (% of Observations with Missing Amounts)	(6) No Bracket Information (% of Observations with Missing Amounts)
Panel A. Income from Dependent Labor						
Sweden	17.75	11.58	21.82	11.92	0.66	87.42
Denmark	10.08	6.27	12.17	20.63	1.25	78.13
Germany	14.94	9.56	24.73	18.84	1.93	79.23
Netherlands	18.28	11.02	31.57	10.61	1.68	87.71
Belgium	10.18	6.32	22.18	19.25	0.53	80.21
France	12.97	9.31	17.38	28.09	2.25	69.66
Switzerland	19.19	9.68	30.33	16.04	2.36	81.60
Austria	9.75	8.09	16.96	10.53	2.63	86.84
Italy	8.69	6.88	17.19	10.75	2.15	87.10
Spain	13.11	7.44	36.35	21.16	5.82	73.02
Greece	9.37	4.33	24.84	9.09	1.01	89.90
Czech Republic	14.66	6.06	28.49	44.30	2.35	53.36
Poland	17.09	13.42	25.86	15.00	1.11	83.89
Panel B. Main Public Pension Income						
Sweden	10.13	1.39	19.11	8.42	3.37	88.22
Denmark	5.74	1.05	13.08	9.09	0.00	90.91
Germany	12.26	2.02	21.92	19.35	3.55	77.10
Netherlands	10.57	3.38	21.00	7.45	1.96	90.59
Belgium	9.61	1.49	20.89	10.80	2.44	86.76
France	11.82	2.30	20.37	22.36	3.63	74.02
Switzerland	9.63	2.36	18.19	9.45	0.00	90.55
Austria	11.14	1.72	16.80	9.66	2.76	87.59
Italy	7.39	1.54	18.00	4.35	0.97	94.69
Spain	9.49	3.02	21.70	12.87	4.68	82.46
Greece	8.40	1.66	20.65	18.11	2.06	79.84
Czech Republic	11.78	0.59	20.66	35.53	2.58	61.89
Poland	13.21	2.89	20.84	1.52	0.00	98.48

Table 5. Non-Responding partners in SHARE

	(1)	(2)	(3)	(4)	(5)	(6)
	Wave 1			Wave 2		
Country	Number	Percentage of the Total Sample	Percentage of Couples with a Non-Responding Partner	Number	Percentage of the Total Sample	Percentage of Couples with a Non-Responding Partner
Sweden	550	15.27	37.75	562	16.99	42.13
Denmark	120	6.57	18.58	240	8.40	21.92
Germany	432	12.56	30.76	351	12.02	28.90
Netherlands	388	11.52	27.60	507	16.00	38.38
Belgium	367	8.75	22.57	328	9.38	24.44
France	180	5.34	14.71	332	10.06	27.15
Switzerland	140	12.24	32.86	273	15.73	41.68
Austria	251	11.71	35.30	167	11.07	32.81
Italy	561	17.98	43.67	331	9.99	23.86
Spain	670	21.85	54.93	259	10.41	26.14
Greece	229	7.32	20.30	247	7.08	18.88
Czech Republic	-..-	-..-	-..-	225	7.36	20.25
Poland	-..-	-..-	-..-	477	16.20	40.66

Table 6. Erroneous zero and monthly values for yearly labor earnings

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Wave 1				Wave 2			
	Zero values		Monthly values		Zero values		Monthly values	
Country	Number	Percentage of the Total Sample	Number	Percentage of the Total Sample	Number	Percentage of the Total Sample	Number	Percentage of the Total Sample
Sweden	26	0.72	68	1.89	162	4.90	59	1.78
Denmark	25	1.37	39	2.13	15	0.53	22	0.77
Germany	33	0.96	83	2.41	26	0.89	84	2.88
Netherlands	21	0.62	59	1.75	12	0.38	102	3.22
Belgium	14	0.33	183	4.36	223	6.38	109	3.12
France	41	1.22	33	0.98	46	1.39	26	0.79
Switzerland	4	0.35	71	6.21	128	7.38	94	5.42
Austria	41	1.91	71	3.31	43	2.85	4	0.27
Italy	21	0.67	52	1.67	149	4.50	50	1.51
Spain	17	0.55	58	1.89	62	2.49	25	1.01
Greece	23	0.74	83	2.65	177	5.07	44	1.26
Czech Republic	-..-	-..-	-..-	-..-	70	2.29	119	3.90
Poland	-..-	-..-	-..-	-..-	19	0.65	36	1.22

Table 7. Quantiles of yearly labor earnings before and after the correction for erroneous monthly values

Country	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Release 2.3.1						Release 2.4					
	1st quantile	5th quantile	10th quantile	25th quantile	50th quantile	75th quantile	1st quantile	5th quantile	10th quantile	25th quantile	50th quantile	75th quantile
						Panel A. Wave 1						
Sweden	632	2,179	5,120	17,429	26,143	32,962	871	2,941	7,843	18,812	26,143	33,223
Denmark	538	2,285	4,571	26,348	37,237	47,051	807	3,361	7,528	26,886	37,640	47,051
Germany	500	1,300	2,400	6,800	23,500	40,000	700	1,800	3,300	12,000	26,400	40,800
Netherlands	600	1,900	3,400	12,000	25,000	40,000	800	2,800	5,000	14,500	27,000	40,081
Belgium	400	975	1,500	4,250	19,336	30,000	600	1,500	2,603	13,000	25,000	37,176
France	1,068	4,200	7,000	13,000	20,000	32,000	1,500	5,000	7,294	13,500	20,386	32,400
Switzerland	652	1,369	1,825	4,432	22,813	53,448	652	1,955	3,650	11,732	35,894	58,662
Austria	340	800	1,200	2,400	15,000	28,000	600	1,200	2,100	10,000	19,600	30,000
Italy	500	1,200	2,000	9,000	16,000	24,000	500	1,500	5,000	11,000	18,000	25,000
Spain	400	691	900	4,327	10,818	17,000	400	800	1,800	6,400	12,000	18,030
Greece	600	1,500	2,700	8,000	14,000	20,000	600	1,700	3,250	8,400	14,000	21,600
						Panel B. Wave 2						
Sweden	456	1,194	2,715	12,866	19,545	26,059	543	1,520	4,669	13,030	19,545	26,059
Denmark	672	2,416	8,053	18,791	25,503	32,214	672	2,684	10,067	18,791	25,771	32,214
Germany	390	1,100	1,600	3,600	15,000	25,500	450	1,500	2,400	8,000	18,000	30,000
Netherlands	502	1,000	1,600	6,000	18,000	26,000	600	1,800	3,000	10,200	19,265	27,000
Belgium	500	1,000	1,500	5,100	17,472	24,537	600	1,550	4,032	13,440	20,160	28,224
France	795	3,000	6,000	13,200	18,500	28,800	960	4,200	7,700	13,200	19,000	29,000
Switzerland	555	926	1,666	3,702	19,745	42,623	665	1,851	3,702	10,988	29,618	44,427
Austria	340	3,000	5,000	13,000	19,800	26,000	340	3,500	5,000	13,000	20,000	26,000
Italy	500	1,000	1,300	10,000	15,500	19,800	600	3,000	6,500	12,000	15,400	18,906
Spain	400	600	1,100	6,300	13,600	18,000	400	600	1,800	9,000	14,000	20,000
Greece	600	1,000	2,500	8,500	15,000	21,100	600	1,600	5,000	9,754	15,500	24,267
Czech Republic	156	267	355	818	4,195	6,399	178	295	533	2,560	4,621	6,470
Poland	130	208	273	1,248	2,599	4,419	130	211	390	1,560	3,041	4,679

Figure 1A. Lack of mixing across implicates

Figure 1B. Successful mixing across implicates

Figure 2. Gelman-Rubin criterion in a case of fast imputation convergence

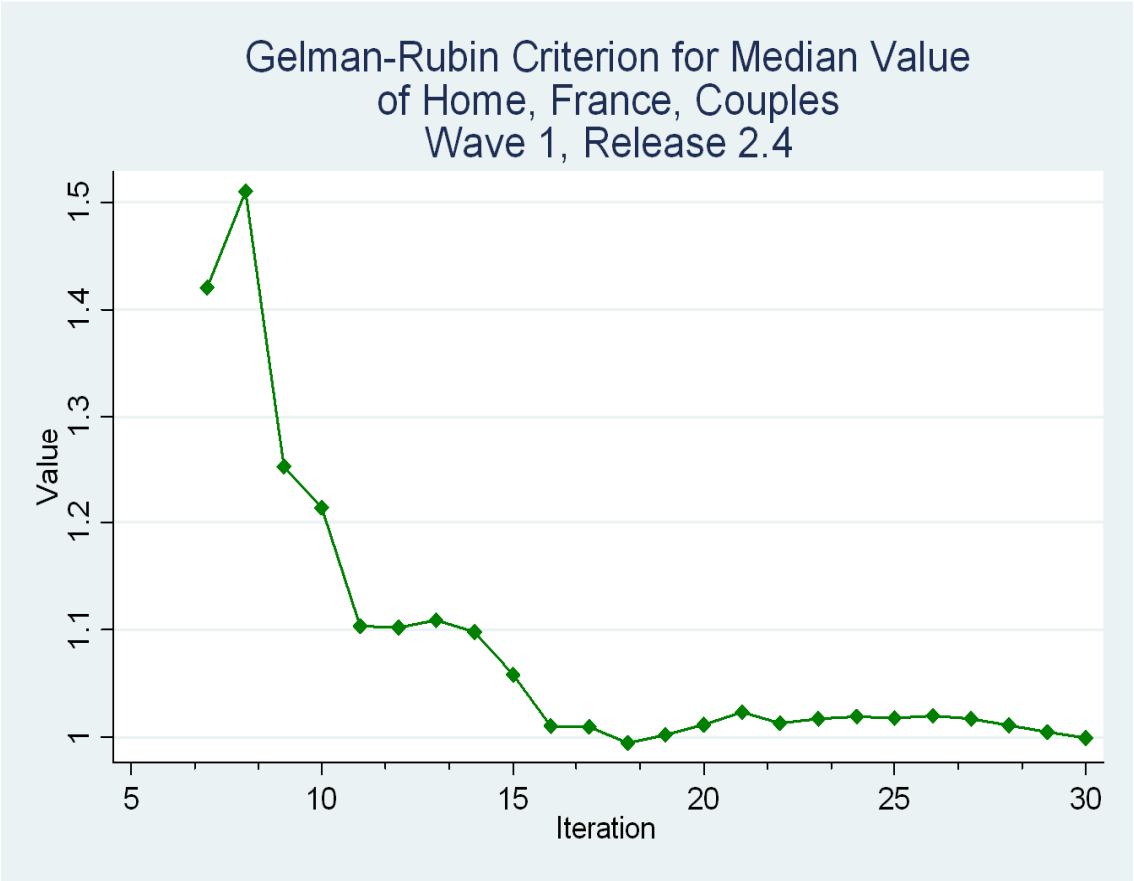


Figure 3. Implicate runs in a case of fast imputation convergence

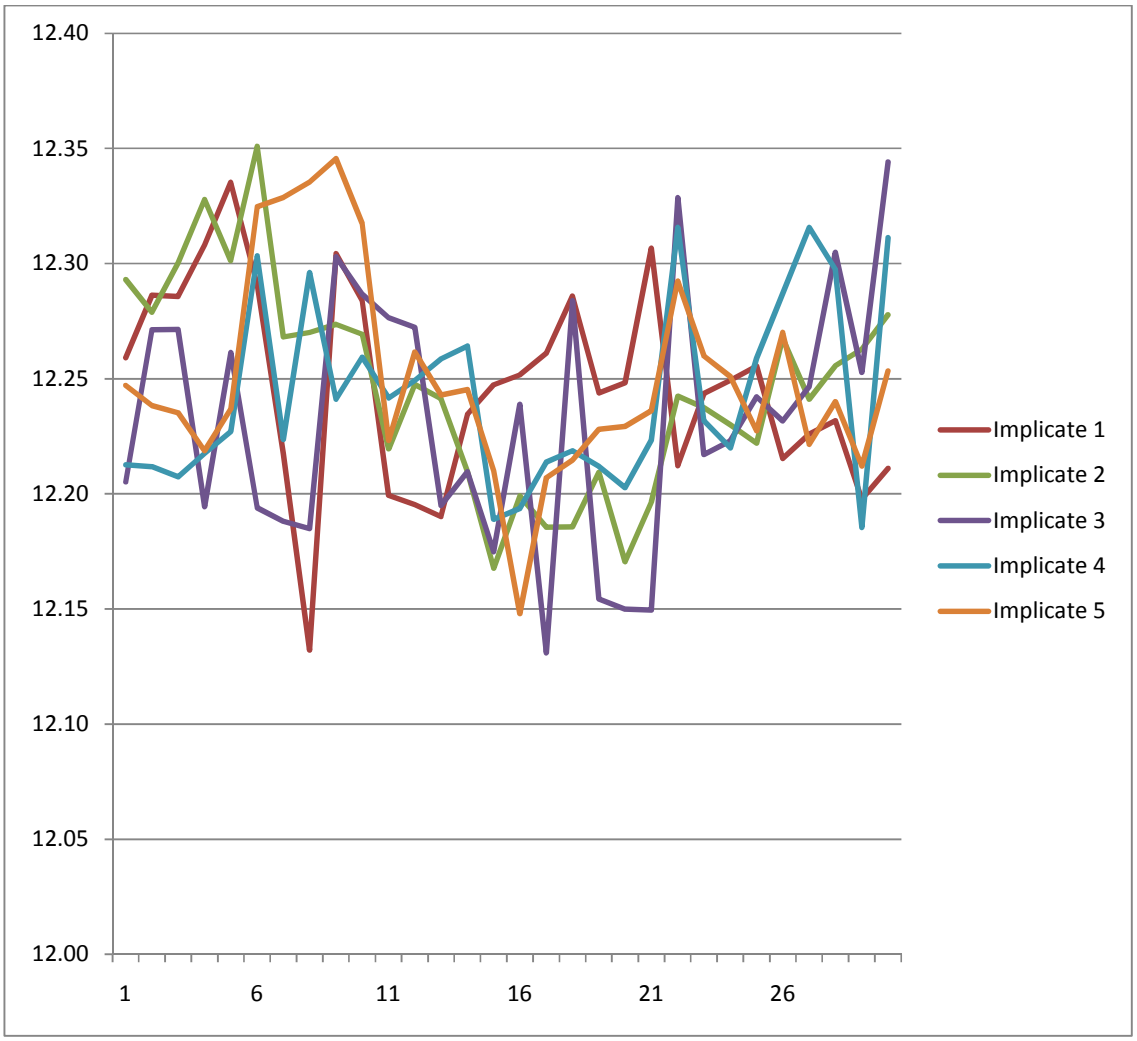


Figure 4. Gelman-Rubin criterion in a case of slow imputation convergence

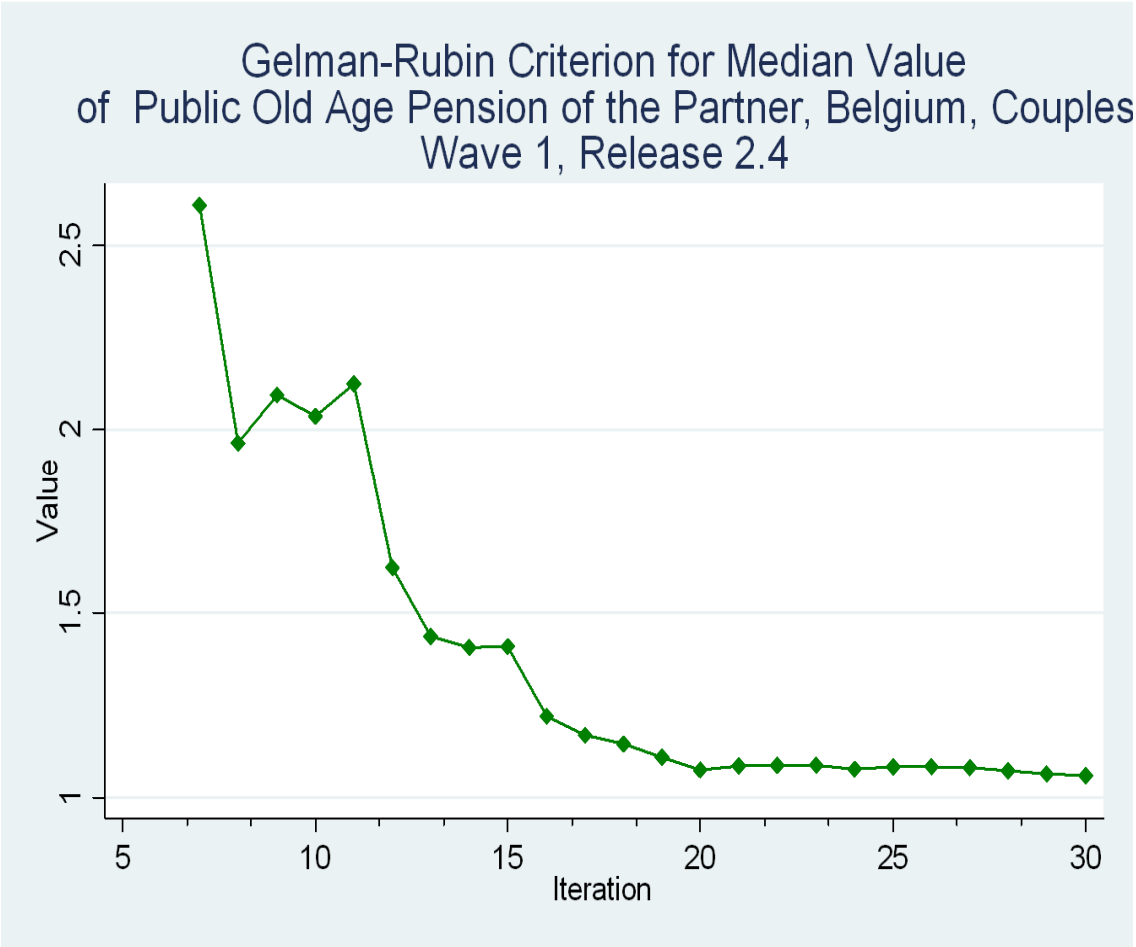


Figure 5. Implicate runs in a case of slow imputation convergence

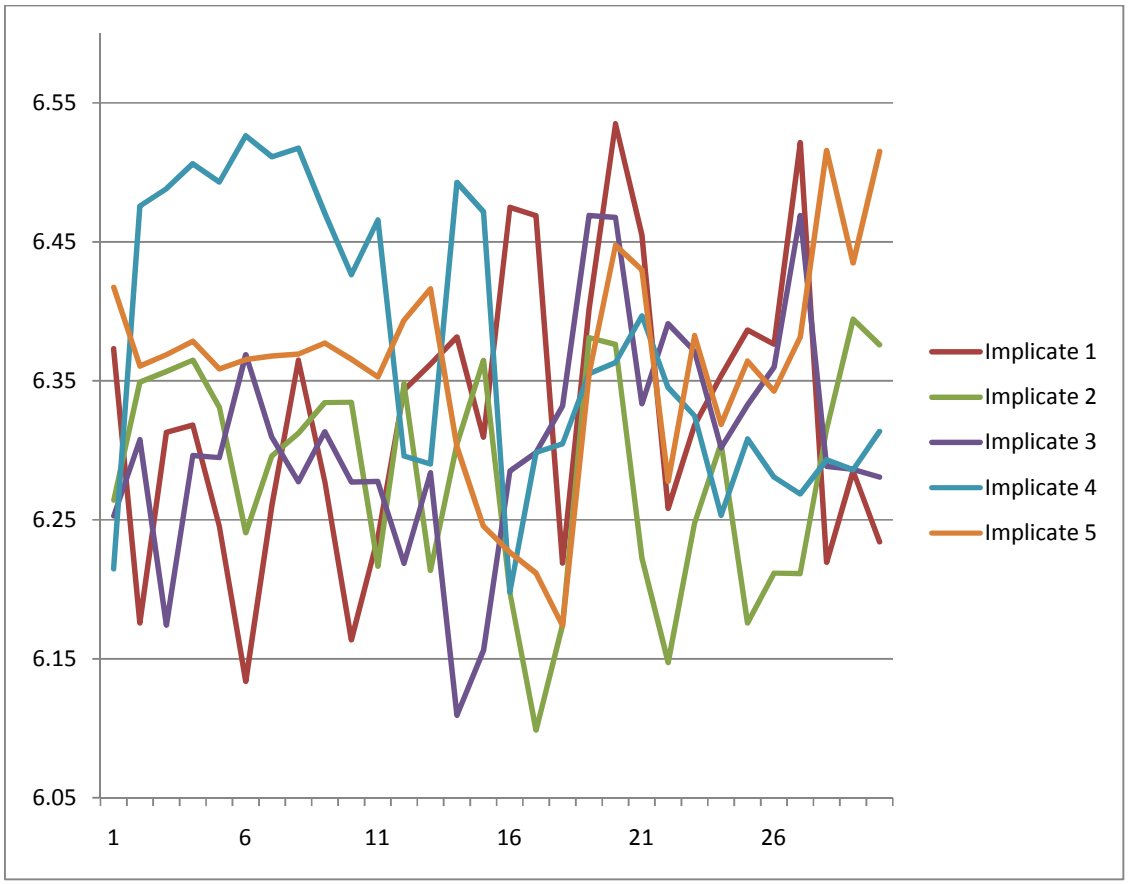


Figure 6. Kernel densities of missing values across iterations

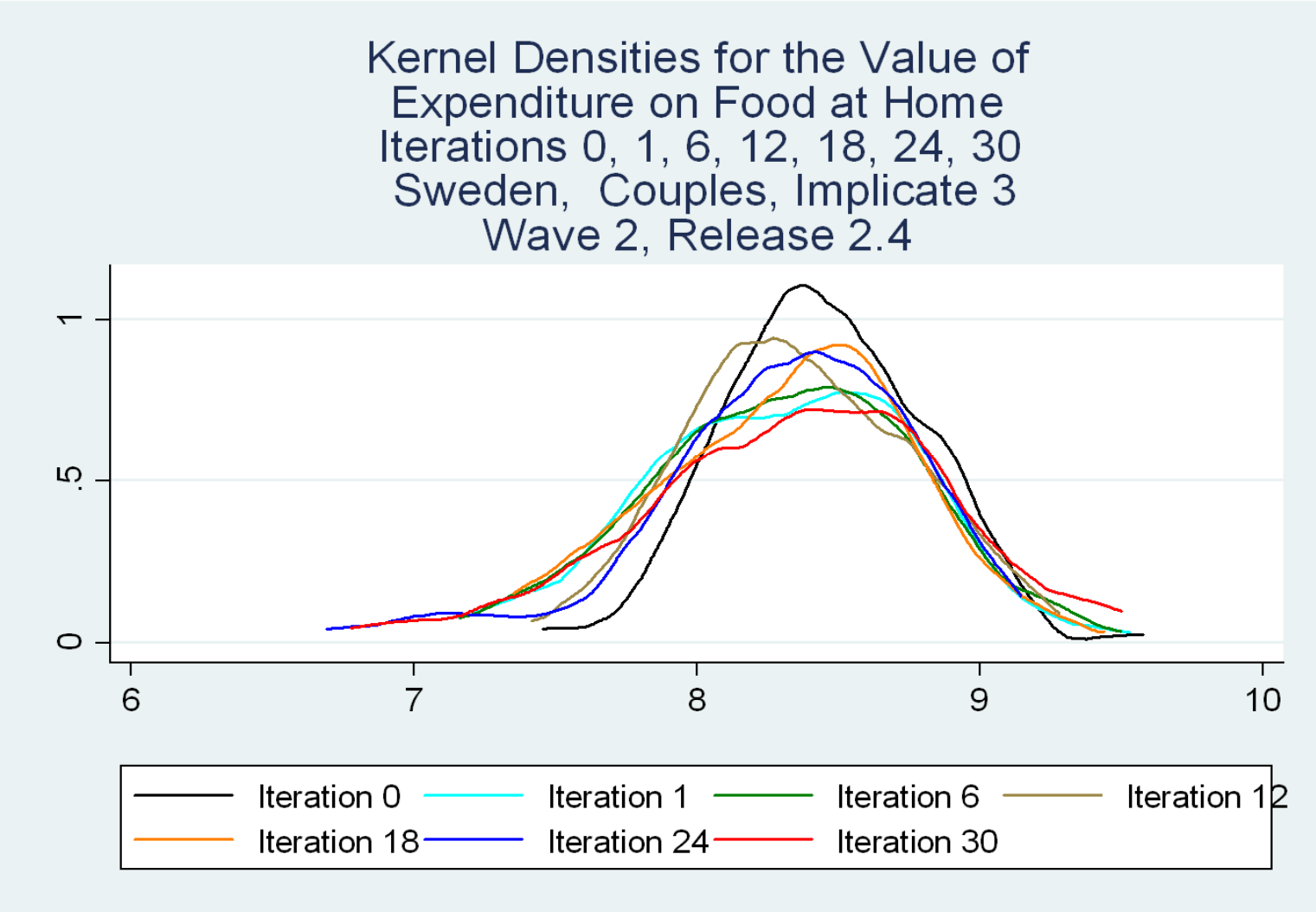


Figure 7. Kernel densities of yearly labor earnings before and after the correction for erroneous monthly values, wave 1

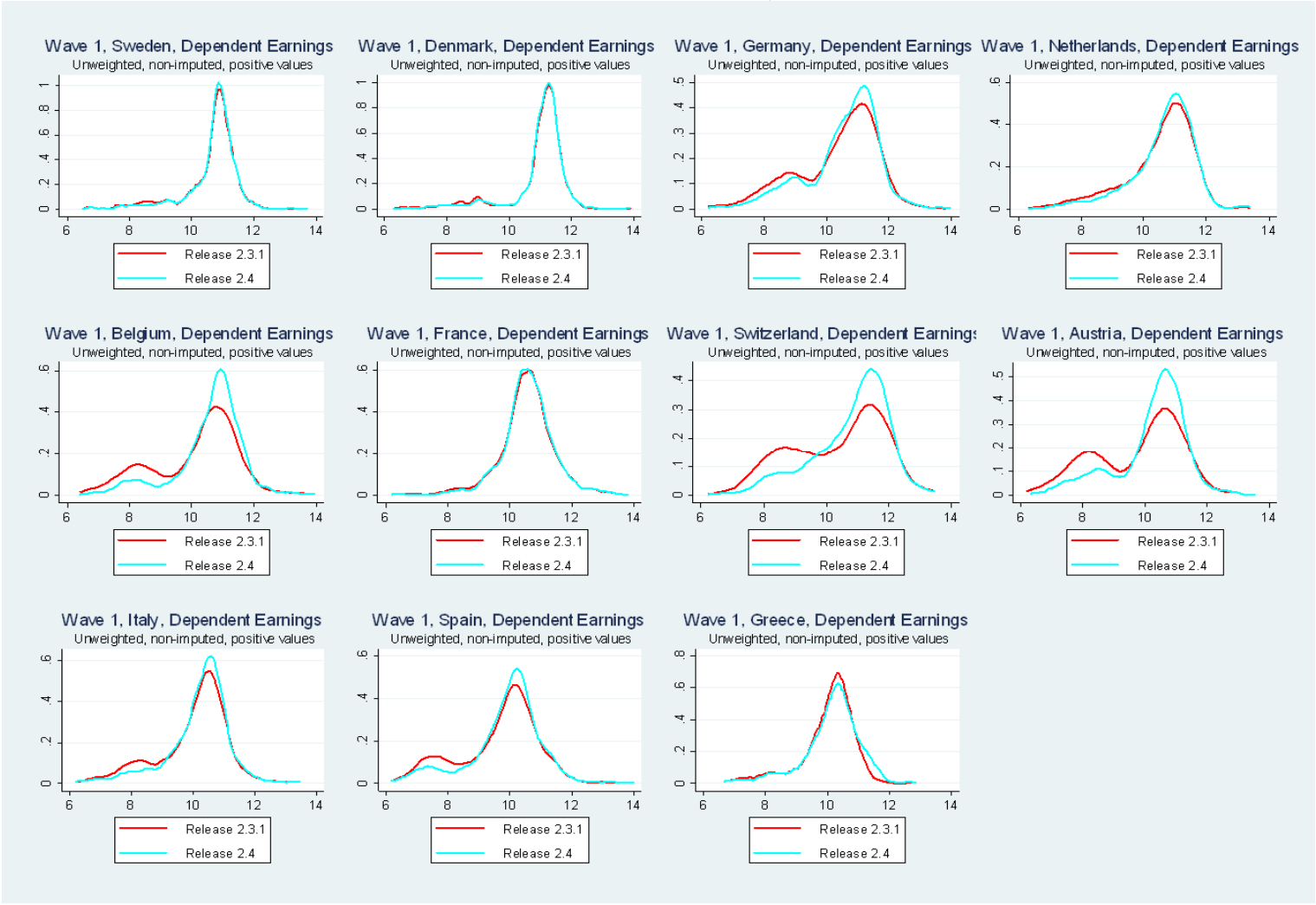
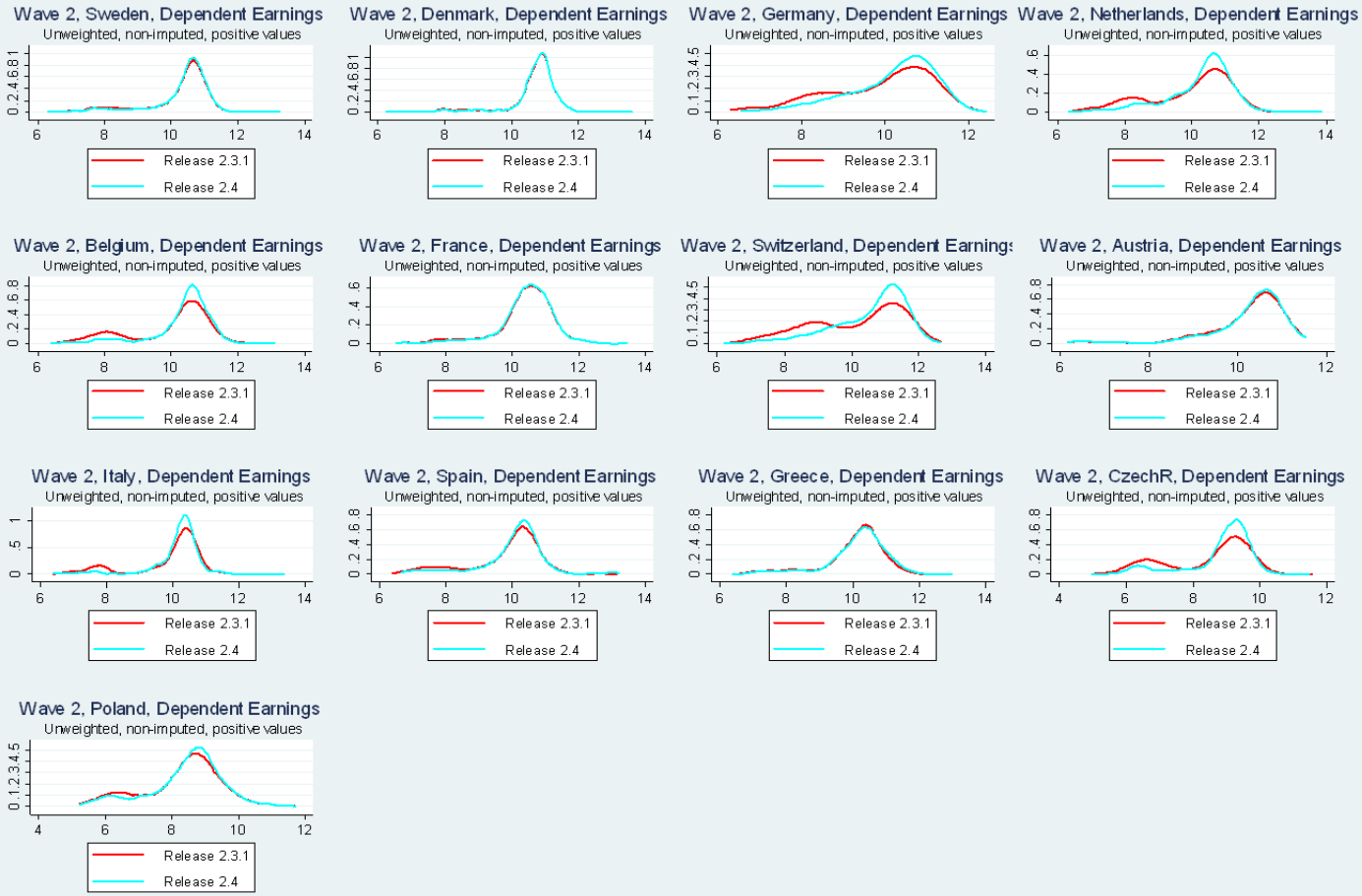


Figure 8. Kernel densities of yearly labor earnings before and after the correction for erroneous monthly values, wave 2



APPENDIX

A.1 Logical constraints imposed during imputation

The logical constraints imposed during the imputation process are as follows:

1. If the household is imputed as having no children, then the number of grandchildren is also imputed to be zero.
2. If the household is imputed to own a home, then the imputed value of the rent payment (and of other rent-related expenses) is set to zero.
3. If the imputation results in lack of home ownership, then the imputed mortgage amount is set to zero.
4. If the first financial transfer given by the household is imputed to be zero, then the second and third ones are also set to zero.
5. If the household is imputed to have bonds, stocks, mutual funds or individual retirement accounts, then it is also imputed to have a bank account.
6. If the household does not have a bank account or does not own any stocks, bonds, and mutual funds, then it does not earn the associated capital incomes.
7. An individual is allowed to have at most three pension items. In other words, if there are already three pension items with positive participation, then any remaining pension items with missing participation will be set to zero.
8. If the household does not own a business, then the owned share of the business is set to zero.

Table A.1. Imputed and generated variables in wave 1

Name	Corresponding Questionnaire Variables	Definition and Comments
		A. Demographics
edu	dn010, dn012	Education, ISCED code
srhealtha	ph003, ph052	Self-reported health, US scale
gali	ph005	Limited in usual activities
numeracy	cf012, cf013, cf014, cf015	Numeracy score
reading	cf001	Self-rated reading skills
adlno	ph048	Number of limitations in ADLs
iadlno	ph049	Number of limitations in IADLs
depress	mh002	Depressed last month
hrooms	ho032	Number of rooms in the main residence
fdistress	co007	Hhd makes ends meet
nchild	ch001	Number of children
n_gchild	ch021	Number of grandchildren
urban	iv009, ho037	Location of the main residence
		B. Individual-level economic variables
ydipv	ep205	Annual gross income from employment previous year
yindv	ep207	Annual gross income from self-employment previous year
pen1v	ep078_1	Monthly public old age pension previous year
pen2v	ep078_2	Monthly public early or pre-retirement pension previous year. In **Sweden**, it refers to invalidity and disability pension
pen3v	ep078_3	Monthly public disability insurance previous year. In **Sweden**, it refers to the survivor pension
pen4v	ep078_4	Monthly public unemployment benefit or insurance previous year. In **Sweden**, it refers to occupational pensions for blue-collar workers in the private sector
pen5v	ep078_5	Monthly public survivor pension from partner previous year. In **Sweden**, it refers to occupational pensions for white-collar workers in the private sector
pen6v	ep078_6	Monthly public invalidity or incapacity pension previous year. In **Sweden**, it refers to occupational pensions for goverment workers
pen7v	ep078_7	Monthly war pension previous year. In **Sweden**, it refers to occupational pension for municipal and local government workers
pen8v	ep078_8	Monthly private (occupational) old age pension previous year. In **Sweden**, it refers to other occupational pension benefit
pen9v	ep078_9	Monthly private (occupational) early retirement pension previous year. In **Sweden**, it refers to unemployment insurance benefits
pen10v	ep078_10	Monthly private (occupational) disability insurance previous year. In **Sweden**, it refers to sickness benefits
pen11v	ep078_11	Monthly private (occupational) survivor pension from partner's job previous year
reg1v	ep094_1	Monthly life insurance payment received previous year
reg2v	ep094_2	Monthly private annuity or private personal pension previous year
reg3v	ep094_3	Monthly private health insurance payment received previous year
reg4v	ep094_4	Monthly alimony received previous year
reg5v	ep094_5	Monthly regular payments from charities received previous year
yltcv	ep086	Monthly long-term care insurance previous year
inpatv	hc045	Out-of-pocket inpatient care expenditure
outpav	hc047	Out-of-pocket outpatient care expenditure
drugsv	hc049	Out-of-pocket expenditure for prescribed medicines
nursv	hc051	Out-of-pocket expenditure for nursing home care, day-care and home care
insurv	hc061	Annual payment for all health insurance contracts
oresv	ho027	Other real estate
yrentv	ho030	Income from rent
mortv	ho015	Mortgage on main residence
baccv	as003	Bank accounts
ybaccv	as005	Interest income from bank accounts
bondv	as007	Government and corporate bonds
ybondv	as009	Interest income from bonds
stocv	as011	Stocks/shares
ystocv	as015	Dividends from stocks/shares
mutfv	as017	Mutual funds
ymutfv	as058	Interest and dividend income from mutual funds
irav	as021, as024	Individual retirement accounts
contv	as027	Contractual savings for housing
linsv	as030	Whole life insurance
gbusv	as042	Total value of (partly) owned business
sbusv	as044	Percentage share of ownership in the business (in percentage points)
ownb	=gbusv*(sbusv/100)	Value of own share of the business
carv	as051	Cars
liabv	as055	Debts (non-mortgage)
ftgiv1v	ft004_1	First financial transfer given
ftgiv2v	ft004_2	Second financial transfer given
ftgiv3v	ft004_3	Third financial transfer given
ftrec1v	ft011_1	First financial transfer received
ftrec2v	ft011_2	Second financial transfer received
ftrec3v	ft011_3	Third financial transfer received

Table A.1 (continued). Imputed and generated variables in wave 1

Name	Corresponding Questionnaire Variables	Definition and Comments
		C. Household-level economic variables
yohmv	hh002	Annual other hhd members' gross income previous year
yohbv	hh011	Annual other hhd members' gross income from other sources previous year
homev	ho024	Hhd main residence
fahcv	co002	Hhd monthly expenditure on food at home
fohcv	co003	Hhd monthly expenditure on food outside the home
telcv	co004	Hhd monthly telephnone expenditure
rentcv	ho005	Hhd monthly rent paid
ocscv	ho008	Hhd monthly other rent-related expenditures
		D. Individual-level Generated Variables
annpen1v		Annual value of pen1v in the previous year
annpen2v		Annual value of pen2v in the previous year
annpen3v		Annual value of pen3v in the previous year
annpen4v		Annual value of pen4v in the previous year
annpen5v		Annual value of pen5v in the previous year
annpen6v		Annual value of pen6v in the previous year
annpen7v		Annual value of pen7v in the previous year
annpen8v		Annual value of pen8v in the previous year
annpen9v		Annual value of pen9v in the previous year
annpen10v		Annual value of pen10v in the previous year
annpen11v		Annual value of pen11v in the previous year
annreg1v		Annual value of reg1v in the previous year
annreg2v		Annual value of reg2v in the previous year
annreg3v		Annual value of reg3v in the previous year
annreg4v		Annual value of reg4v in the previous year
annreg5v		Annual value of reg5v in the previous year
		E. Household-level Generated Variables
hmortv		HHd mortgage on main residence
horesv		HHd other real estate
hbaccv		Hhd bank accounts
hbondv		Hhd government and corporate bonds
hstocv		Hhd stocks/shares
hmutfv		Hhd mutual funds
hirav		Hhd individual retirement accounts
hcontv		Hhd contractual savings for housing
hlinsv		Hhd whole life insurance
hownbv		Hhd value of own share of businesses
hcarv		Hhd cars
hliabv		Hhd debts (non-mortgage)
hybaccv		Hhd interest income from bank accounts
hybondv		Hhd interest income from bonds
hystocv		Hhd dividends from stocks/shares
hymutfv		Hhd interest and dividend income from mutual funds
hyrentv		Hhd income from rent
hrav		Hhd real assets net of any debts on them.Their value is equal to the sum of **homev**, **horesv**, **hownbv**, **hcarv** minus **hmortv**
hgfinv		Hhd gross financial assets.Their value is equal to the sum of **hbaccv**, **hbondv**, **hstocv**, **hmutfv**, **hirav**, **hcontv**, and **hlinsv**
hnfinv		Hhd net financial assets. Their value is equal to **hgfinv** minus **hliabv**
hnetwv		Hhd net worth. Its value is equal to the sum of **hrav** and **hnfinv**
hgtincv		Hhd total gross income. Its value is equal to the sum over all household members of the individual-level values of **ydipv, yindv**, **annpen1v** – **annpen11v**, **annreg1v** – **annreg5v**, 12 times **yltcv, ybaccv, ybondv**, **ystocv**, **ymutfv**, **yrentv**. To this sum one has to add the sum of the values of the household-level variables **yohmv** and **yohbv**.

Table A.2. Imputed and generated variables in wave 2

Name	Corresponding Questionnaire Variables	Definition and Comments
		A. Demographics
edu	dn010, dn012	Education, ISCED code
srhealtha	ph003	Self-reported health, US scale
gali	ph005	Limited in usual activities
numeracy	cf012, cf013, cf014, cf015	Numeracy score
reading	cf001	Self-rated reading skills (only for refresher sample)
adlno	ph048	Number of limitations in ADLs
iadlno	ph049	Number of limitations in IADLs
depress	mh002	Depressed last month
hrooms	ho032	Number of rooms in the main residence
fdistress	co007	Hhd makes ends meet
nchild	ch001	Number of children
n_gchild	ch021	Number of grandchildren
urban	iv009, ho037	Location of the main residence
riskpref	as068	Risk preferences
		B. Individual-level economic variables
ydipv	ep205	Annual net income from employment, previous year
yindv	ep207	Annual net income from self-employment, previous year
pen1v	ep078_1	Monthly public old age pension, previous year
pen2v	ep078_3	Monthly public early or pre-retirement pension, previous year. In **Sweden**, it refers to invalidity and disability pension
pen3v	ep078_4	Monthly main public disability insurance pension, or sickness benefits, previous year. In **Sweden**, it refers to the survivor pension
pen4v	ep078_6	Monthly public unemployment benefit or insurance, previous year. In **Sweden**, it refers to occupational pensions for blue-collar workers in the private sector
pen5v	ep078_7	Monthly public survivor pension from partner, previous year. In **Sweden**, it refers to occupational pensions for white-collar workers in the private sector
pen7v	ep078_9	Monthly war pension, previous year. In **Sweden**, it refers to occupational pension for workers in municipalities, in counties or in the government
pen8v	ep324_1	Monthly private (occupational) old age pension, previous year
pen9v	ep324_4	Monthly private (occupational) early retirement pension, previous year. In **Sweden**, it refers to unemployment insurance benefits
pen10v	ep324_5	Monthly private (occupational) disability insurance, previous year. In **Sweden**, it refers to sickness benefits
pen11v	ep324_6	Monthly private (occupational) survivor pension from partner's job, previous year
pen12v	ep078_2	Monthly public old age supplementary pension or public old age second pension, previous year
pen13v	ep078_5	Monthly secondary public disability insurance pension, or sickness benefits, previous year
pen14v	ep078_8	Monthly secondary public survivor pension from spouse or partner, previous year
pen15v	ep324_2	Monthly occupational old age pension from a second job, previous year
pen16v	ep324_3	Monthly occupational old age pension from a third job, previous year
pen17v	ep324_5	(**only in Sweden**) - Monthly private (occupational) disability insurance, previous year
pultv	ep078_10	Monthly public long-term insurance payments, previous year
reg1v	ep094_1	Monthly life insurance payment received, previous year
reg2v	ep094_2	Monthly private annuity or private personal pension, previous year
reg3v	ep094_2	(**only in Greece**) Monthly private health insurance payment received, previous year
reg4v	ep094_3	Monthly alimony received, previous year
reg5v	ep094_4	Monthly regular payments from charities received, previous year
prltv	ep094_5	Monthly private long-term care insurance payments, previous year
inpatv	hc045	Out-of-pocket inpatient care expenditure, annual, previous year
outpav	hc047	Out-of-pocket outpatient care expenditure, annual, previous year
drugsv	hc049	Out-of-pocket expenditure for prescribed medicines, annual, previous year
nursv	hc051	Out-of-pocket expenditure for nursing home care, day-care and home care, annual, previous
oresv	ho027	Other real estate
yrentv	ho030	Income from rent
mortv	ho015	Mortgage on main residence
ftgiv1v	ft004_1	First financial transfer given
ftgiv2v	ft004_2	Second financial transfer given
ftgiv3v	ft004_3	Third financial transfer given
ftrec1v	ft011_1	First financial transfer received
ftrec2v	ft011_2	Second financial transfer received
ftrec3v	ft011_3	Third financial transfer received

Table A.2 (continued). Imputed and generated variables in wave 2

Name	Corresponding Questionnaire Variables	Definition and Comments
		C. Household-level economic variables
yohmv	hh002	Annual other hhd members' net income previous year
yohbv	hh011	Annual other hhd members' net income from other sources previous year
homev	ho024	Hhd main residence
hbaccv	as003	Hhd bank accounts
hbondv	as007	Hhd government and corporate bonds
hstocv	as011	Hhd stocks/shares
hmutfv	as017	Hhd mutual funds
hirav	as021, as024	Hhd individual retirement accounts
hcontv	as027	Hhd contractual savings for housing
hlinsv	as030	Hhd whole life insurance
hownbv	as042, as044	Hhd value of own share of businesses
hcarv	as051	Hhd cars
hliabv	as055	Hhd debts (non-mortgage)
hybaccv	as005	Hhd interest income from bank accounts
hybondv	as009	Hhd interest income from bonds
hystocv	as015	Hhd dividends from stocks/shares
hymutfv	as058	Hhd interest and dividend income from mutual funds
fahcv	co002	Hhd monthly expenditure on food at home
fohcv	co003	Hhd monthly expenditure on food outside the home
telcv	co004	Hhd monthly telephnone expenditure
hprcv	co011	Hhd monthly home production of food
rentcv	ho005	Hhd monthly rent paid
ocscv	ho008	Hhd monthly other rent-related expenditures
		D. Individual-level Generated Variables
annpen1v		Annual value of pen1v in the previous year
annpen2v		Annual value of pen2v in the previous year
annpen3v		Annual value of pen3v in the previous year
annpen4v		Annual value of pen4v in the previous year
annpen5v		Annual value of pen5v in the previous year
annpen7v		Annual value of pen7v in the previous year
annpen8v		Annual value of pen8v in the previous year
annpen9v		Annual value of pen9v in the previous year
annpen10v		Annual value of pen10v in the previous year
annpen11v		Annual value of pen11v in the previous year
annpen12v		Annual value of pen12v in the previous year
annpen13v		Annual value of pen13v in the previous year
annpen14v		Annual value of pen14v in the previous year
annpen15v		Annual value of pen15v in the previous year
annpen16v		Annual value of pen16v in the previous year
annpen17v		Annual value of pen17v in the previous year (only exists in **Sweden**)
annpultv		Annual value of pultv in the previous year
annreg1v		Annual value of reg1v in the previous year
annreg2v		Annual value of reg2v in the previous year
annreg3v		Annual value of reg3v in the previous year
annreg4v		Annual value of reg4v in the previous year
annreg5v		Annual value of reg5v in the previous year
annprltv		Annual value of prltv in the previous year
		E. Household-level Generated Variables
hmortv		HHd mortgage on main residence
horesv		HHd other real estate
hyrentv		Hhd income from rent
hrav		Hhd real assets net of any debts on them.Their value is equal to the sum of **homev**, **horesv**, **hownbv**, **hcarv** minus **hmortv**
hgfinv		Hhd gross financial assets.Their value is equal to the sum of **hbaccv**, **hbondv**, **hstocv**, **hmutfv**, **hirav**, **hcontv**, and **hlinsv**
hnfinv		Hhd net financial assets. Their value is equal to **hgfinv** minus **hliabv**
hnetwv		Hhd net worth. Its value is equal to the sum of **hrav** and **hnfinv**
hgtincv		Hhd total gross income. Its value is equal to the sum over all household members of the individual-level values of **ydipv, yindv**, **annpen1v** – **annpen5v**, **annpen7v – annpen16v**, **annpultv**, **annprltv**, **annreg1v** – **annreg5v**, **yrentv**. To this sum one has to add the sum of the values of the household-level variables **yohmv**, **yohbv**, **hybaccv**, **hybondv**, **hystocv**, and **hymutfv**.